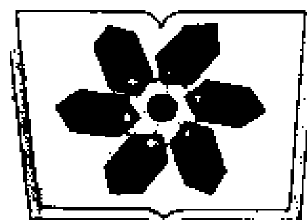
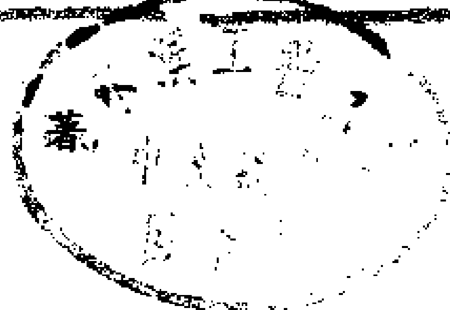


中国科学院科学出版基金资助出版

成分数据统计分析引论



张尧庭 著

科学出版社

2000

内 容 简 介

成分数据统计分析的主要内容是以成分数据为目标的统计理论与方法，其基础与多种分布，如逻辑正态分布族，狄氏分布族等有关．本书在此基础上介绍了成分数据统计分析的理论与分析方法，以及这一方向在国内外的最新成果．本书每章末附有习题，以便读者更好地理解本书内容．

本书可供应用统计工作者、科研人员及大学有关专业高年级学生、研究生、教师阅读．

图书在版编目(CIP)数据

成分数据统计分析引论/张尧庭著．-北京:科学出版社,2000
ISBN 7-03-008264-8

Ⅰ．成…　Ⅱ．张…　Ⅲ．数理统计　Ⅳ．O212.1

中国版本图书馆 CIP 数据核字（2000）第 01520 号

科 学 出 版 社 出版
北京东黄城根北街 16 号
邮政编码:100717
新 蕾 印 刷 厂 印刷

科学出版社发行　各地新华书店经销

*

2000 年 8 月第 一 版　开本：850×1168　1/32
2000 年 8 月第一次印刷　印张：5 3/8
印数：1—2 000　字数：138 000

定价：14.00 元

（如有印装质量问题，我社负责调换（新欣））

前　言

成分数据的统计问题早就提出了，早在1897年，K·皮尔逊就指出解决这些问题是很困难的．直到1986年，才有第一本书专门论述这一类问题，作者J. Aitchison也因在这一方向研究的特殊贡献，荣获英国皇家统计学会1988年的研究奖章．此书的中译本（《成分数据的统计分析》）在1990年由中国地质大学出版社出版．艾奇逊在书的前言中阐明了他写这本书的目的是："……在于清晰而完整地介绍近年来专为成分数据而设计的一套新的统计方法．……强调提供有效而可行的方法，而不是进行详细的证明和推导．"这样的写法对于想进一步了解理论的人，是会感到不满足的．新的方法主要是指基于加性逻辑正态分布的统计分析方法．

成分数据的统计分析与单形上的分布有密切的联系，逻辑正态、狄氏分布等等都是单形上的分布族．只有把单形上的分布类型作一些合理的分类，才可能把成分数据的统计分析放在一个坚实的基础上，才有可能提供各种恰当的分析方法．从理论上看，这样也比较系统和自然．

从成分数据的实际背景来看，加性、乘性逻辑正态分布仅仅反映了一种类型的分布．被艾奇逊认为不够丰富的狄氏分布及其推广，事实上也是重要的．本书发展了这一方面的统计分析方法，对理论和实际都提供了新的工具．

有了单形上的分布，混料试验的设计就可以有一个比较恰当的理论模型．一方面既可以说明为什么混料试验中添加对数项往往会有效，另一方面也推动了设计的改进和分析方法的多样化，这些内容在艾奇逊书中没有展开，而在本书中有一些反映．本书尽量不和艾奇逊书的内容重复，因为我的观点和他有些不同．我

认为，直到现在，真正的成分数据的分析仍然是非常困难的. 本书一个重要的目的是显示它的困难，但我没有办法解决，而是希望有更多的人来从事这一方面的研究.

本书断断续续写了几乎 7 年，因为这几年有不少其他的事情，迫使我中断下来，所以本书的一、二章和三、四章有明显的差别，我自己也感到不满意，由于能力有限，只能这样了. 书中有一些内容是没有发表过的成果，但我感到没有做完、做好，只能重印时再说了.

本书的前两章，邹国华同志看了之后，提出了不少好的意见，并改正了不少错误，在这里我表示衷心的感谢；另外书中也汇集了章栋恩同志近年来的工作，有些即将发表，我得到他的允许，也反映到本书中.

本书的出版得到了中国科学院出版基金的支持，我深表感谢. 希望这本书既能填补我国这一方面的空白，也希望为将来的研究提供条件. 我还要感谢多年来一直合作很好的毕颖同志，她为本书的出版和编辑付出了大量的精力.

希望得到读者的批评和帮助，以便改正由于我水平有限所出现的不足.

张尧庭

1999 年 11 月

目　录

第一章　准备知识

§1. n 维欧氏空间与单形

我们在这一节引入本书常用的一些记号，并列举一些重要的概念，读者可以从一般的线性代数的教材中查阅有关的说明．我们假定读者已具备线性代数、数理统计的基本知识，并对多元统计分析的方法有一些了解．

用 R_n 表示实数域上的 n 维线性空间，R_n 中的向量用小写字母表示，如 $a,b,c,\cdots,x,y,z$ 等．向量的分量以相应的带脚标的字母表示，如

$$a=\begin{pmatrix}a_1\\ \vdots\\ a_n\end{pmatrix},\qquad x=\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix}.$$

通常矩阵用大写字母表示，如 $A,B,C,\cdots,X,Y,Z$ 等．它的大小列举在字母下面，如 $\underset{n\times m}{A}$ 表示 A 是 n 行 m 列的矩阵，从上下文可以确定其大小时，就不再列举，A 中的元素用双脚标表示，如

$$A=(a_{ij}),\qquad X=(x_{ij}).$$

矩阵的转置用“$'$”表示，A' 表示将 A 转置后的矩阵．我们总是把向量也看成矩阵，当

$$a=\begin{pmatrix}a_1\\ \vdots\\ a_n\end{pmatrix}$$

时，a 的转置 $a'=(a_1,a_2,\cdots,a_n)$．于是向量 a 与 b 的内积可以用 $a'b$ 写出，即

$$a'b = (a_1, \cdots, a_n)\begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \sum_{i=1}^{n} a_i b_i,$$

在 R_n 中引入内积后，R_n 就是一个 n 维的欧氏空间. 对于矩阵，有时也用它的行向量或列向量表示，通常都用一个脚标，带括号表示行指标，不带括号表示列指标，即

$$\underset{n\times m}{A} = (a_{ij}) = \begin{pmatrix} a'_{(1)} \\ \vdots \\ a'_{(n)} \end{pmatrix} = (a_1 \quad a_2 \quad \cdots \quad a_m),$$

易见

$$a'_{(i)} = (a_{i1}, a_{i2}, \cdots, a_{im}), \quad a'_i = (a_{1i}, a_{2i}, \cdots, a_{ni}).$$

我们用 $\mathbb{1}_n$ 表示元素全为 1 的 n 维向量，当维数由上下文确定时，用 $\mathbb{1}$ 表示，于是有

$$\mathbb{1}'a = a'\mathbb{1} = (a_1, \cdots, a_n)\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \sum_{i=1}^{n} a_i,$$

在统计中常见的一组数据 $a_1, \cdots, a_n$ 的均值和方差均可用向量、矩阵的运算表示出来，因为

$$\frac{1}{n}\sum_{i=1}^{n} a_i = \frac{1}{n}\mathbb{1}'a = \frac{1}{n}a'\mathbb{1} = \bar{a},$$

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(a_i - \bar{a})^2 &= \frac{1}{n}\left(\sum_{i=1}^{n} a_i^2 - n\bar{a}^2\right) \\ &= \frac{1}{n}\left(a'a - n\left(\frac{1}{n}\mathbb{1}'a\right)^2\right) \\ &= \frac{1}{n}a'\left(I_n - \frac{1}{n}\mathbb{1}\mathbb{1}'\right)a, \end{aligned}$$

其中 $a = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$，$I_n$ 就是 $n \times n$ 的单位阵.

当向量 a 的每个分量均大于零时,我们用 $a>0$ 表示,类似的 $a\geqslant 0$ 就表示 a 的分量均为非负的实数.然而矩阵 $A\geqslant 0$ 则表示 A 是非负定的方阵,$A>0$ 表示 A 是正定阵.

向量 $\underset{n\times 1}{x}$ 的一个线性函数 $\sum_{i=1}^{n} a_i x_i$ 可以写成 $a'x$,其中常数向量 a 是由系数 $a_1,\cdots,a_n$ 形成的.于是线性方程组

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = b_1, \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m = b_m. \end{cases}$$

可以写成矩阵或向量的形式:

$$\underset{n\times m}{A}\ \underset{m\times 1}{x} = \underset{n\times 1}{b}$$

或

$$a_1x_1 + a_2x_2 + \cdots + a_mx_m = b,$$

而

$$A = (a_{ij}), \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix},$$

A 也可写成$(a_1 \quad a_2 \quad \cdots \quad a_m)$.

给定矩阵 $\underset{n\times m}{A}=(a_1 \quad a_2 \quad \cdots \quad a_m)$后,A 的列向量$a_1,\cdots,a_m$ 所张成的线性子空间用$R(A)$表示,即

$$R(A) = \{\underset{n\times m}{A}\ \underset{m\times 1}{x} : x \in R_m\}.$$

很明显,$R(A)$是 R_n 中的一个子空间.当 $a'b=0$ 时,我们称 a 与 b 正交,记为 $a\perp b$.若 a 与集合 S 中每一个向量都正交,就记为 $a\perp S$.如果 b 与矩阵 $A=(a_1 \quad a_2 \quad \cdots \quad a_m)$中的每一个列向量都正交,即 $b\perp a_i, i=1,2,\cdots,m$,则 $b\perp R(A)$.与$R(A)$正交的向量形成一个子空间,称为 $R(A)$的正交补空间,记为 $R(A)^{\perp}$,即

$$R(A)^{\perp} = \{x : x \perp R(A)\}.$$

很明显,$R(A)^{\perp}$也可以用另一个形式表示,即

$$R(A)^{\perp} = \{x : x \perp a_i, i = 1,2,\cdots,m\}$$

$$= \{x : a_i'x = 0, i = 1,2,\cdots,m\}$$
$$= \{x : A'x = 0\}.$$

考虑一个特殊的正交补空间,它在今后的统计分析中起重要的作用.若 $a'\mathbb{1}=0$,则称 $a'x$ 是向量 x 的一个对比,a 称为这个对比的系数向量.如

$$x_1 - x_2, \quad x_1 - 2x_2 + x_n,$$
$$x_1 - \frac{1}{n-1}(x_2 + \cdots + x_n), \cdots$$

都是 x 的对比.很明显,$R(\mathbb{1})^{\perp}$就是对比系数向量组成的子空间,它就是方程

$$\mathbb{1}'x = 0$$

的全部的解.

为了方便,今后用一些记号代表 R_n 中的一些特定的集合.这些集合是:

$$R_n^+ = \{x : x \in R_n, x_i > 0, i = 1,2,\cdots,n\},$$
$$\overline{R}_n^+ = \{x : x \in R_n, x_i \geqslant 0, i = 1,2,\cdots,n\},$$
$$S_n = \{x : x \in R_n^+, \mathbb{1}'x = 1\},$$
$$\overline{S}_n = \{x : x \in \overline{R}_n^+, \mathbb{1}'x = 1\},$$
$$D_n = \{x : x \in R_n^+, \mathbb{1}'x < 1\},$$
$$\overline{D}_n = \{x : x \in \overline{R}_n^+, \mathbb{1}'x \leqslant 1\},$$

其中集合$\overline{S}_n$ 通常称为 n 维空间中的单形,或简称为单形(有的书上称为单纯形),S_n 是单形 $\overline{S}_n$ 的内部.

成分数据,也就是由百分比组成的数据,因此成分数据的取值范围就是单形$\overline{S}_n$,由于讨论时在数学上更便于处理,我们常常先讨论 S_n 内取值的情况,所以 $S_n, \overline{S}_n$ 是我们今后经常会遇到的集合.

如果 $a_i \geqslant 0, i = 1,2,\cdots,k$,且 $\sum_{i=1}^{k} a_i = 1$,也即 $a = \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix} \in \overline{S}_k$,

$b_1, b_2, \cdots, b_k$ 均为 R_n 中的向量，则称

$$a_1 b_1 + a_2 b_2 + \cdots + a_k b_k = (b_1\ b_2 \cdots\ b_k) a$$

是 $b_1, \cdots, b_k$ 的凸线性组合. 用矩阵 $\underset{n \times k}{B} = (b_1 \cdots\ b_k)$ 来表示，就是：只要 $a \in \overline{S}_k$，Ba 就是 B 的列向量的一个凸线性组合，集合

$$\{Ba : a \in \overline{S}_k\}$$

就是向量 $b_1, b_2, \cdots, b_k$ 所生成的凸包. R_n 中一个集合对凸线性组合是封闭的，也即其中任意有限个向量的凸线性组合仍属于这个集合，则称这个集合是凸集. 很明显，$R_n^+, \overline{R}_n^+, S_n, \overline{S}_n, D_n, \overline{D}_n$ 都是凸集，$b_1, \cdots, b_k$ 生成的凸包 $\{Ba : a \in \overline{S}_k\}$ 是包含 $b_1, \cdots, b_k$ 的最小的凸集.

§2. 基和成分

给定一个向量 $\omega \in \overline{R}_{n+1}, \omega \neq 0, \omega' = (\omega_0, \omega_1, \cdots, \omega_n)$，我们就称 ω 是一个基向量，令

$$\begin{cases} t = \sum_{i=0}^{n} \omega_i, \\ x_i = \dfrac{\omega_i}{t}, \end{cases} \qquad i = 0, 1, 2, \cdots, n. \tag{2.1}$$

则称 t 是 ω 相应的总量，$x' = (x_0, x_1, \cdots, x_n)$ 称为 ω 相应的成分，当 ω 明确时，简称为总量、成分. 易见 ω, t, x 有下列关系：

$$\begin{cases} \omega = tx, x = \omega t^{-1} (\omega \neq 0), \\ t \geqslant 0, x \geqslant 0, \mathbb{1}'x = 1 (\omega \neq 0). \end{cases} \tag{2.2}$$

为了今后讨论方便，以下无特殊声明时，我们总假定 $\omega \in R_{n+1}^+$，于是相应有

$$\begin{cases} \omega = tx, x = \omega t^{-1}, \\ t > 0, x > 0, \mathbb{1}'x = 1. \end{cases} \tag{2.3}$$

此时，我们对向量 ω 或 x 的各个分量取对数是有意义的，取对数后形成的向量分别用 $\ln\omega$ 和 $\ln x$ 表示，即

$$\ln\omega = \begin{pmatrix} \ln\omega_0 \\ \ln\omega_1 \\ \vdots \\ \ln\omega_n \end{pmatrix}, \quad \ln x = \begin{pmatrix} \ln x_0 \\ \ln x_1 \\ \vdots \\ \ln x_n \end{pmatrix}.$$

我们称 $\ln\omega$ 的一个对比 $a'\ln\omega$ 是 ω 的一个对数对比.对成分 x 也是一样,当 $\mathbb{1}'a=0$ 时,$a'\ln x$ 称为 x 的一个对数对比.

如果把 $x\in S_{n+1}$ 看成一个基向量,则 x 相应的成分还是它自己.因此基向量 ω 和相应的成分 x 可以看成是 ω 在 S_{n+1} 上的"投影",$x\in S_{n+1}$ 时它的"投影"还是自己,不同的 ω 可以有相同的"投影".把这一概念更一般化,就引出形状和大小的抽象概念.

定义 2.1 假定 $\omega\in R_{n+1}^{+}$,即 ω 是一个基向量,如果 $G(\omega)$ 是一个 ω 的正值函数,且有

$$G(c\omega) = cG(\omega), \tag{2.4}$$

对一切 $c>0$ 成立,则称 $G(\omega)$ 是 ω 的大小(Size),令

$$z_G(\omega) = \omega/G(\omega), \tag{2.5}$$

称 $z_G(\omega)$ 是 $G(\omega)$ 这个大小相应的形状(shape).

从定义 2.1 可以看出,总量 t 是一个 ω 的大小,它相应的形状就是成分 x.很明显,ω 的大小可以列举很多,如

$$\left(\sum_{i=0}^{n}\omega_i^2\right)^{\frac{1}{2}}, \quad \omega_0, \quad \max_{0\leqslant i\leqslant n}\omega_i$$

等都是,它们各自相应的形状是不同的.然而形状向量之间却有如下的重要关系.

定理 2.1 假定 $\omega\in R_{n+1}^{+}$,G_1,G_2 是 ω 的两个大小,z_1,z_2 是 G_1,G_2 分别相应的形状,则有

$$z_1(\omega) = z_2(\omega)/G_1(z_2(\omega)). \tag{2.6}$$

证明
$$\begin{aligned} z_1(\omega) &= \omega/G_1(\omega) \\ &= \left[\frac{\omega}{G_2(\omega)}\right]\Big/\left[\frac{G_1(\omega)}{G_2(\omega)}\right] \\ &= z_2(\omega)/G_1(\omega/G_2(\omega)) \\ &= z_2(\omega)/G_1(z_2(\omega)), \end{aligned}$$

这就告诉我们,形状向量是可以相互表示的.

现在引入子成分的概念.成分的一部分并不再是一个成分向量,例如将成分向量 x 分为二段,即

$$x = \begin{pmatrix} x_{(1)} \\ x_{(2)} \end{pmatrix},$$

$x_{(i)}$是 $n_i \times 1$ 的向量,$i=1,2$.由于 $x_{(i)}>0$,必然有 $0<\mathbb{1}'x_{(i)}<1$,$i=1,2$,它们各自的分量之和一定小于 1,不能是一个成分向量.子成分是把成分分解为若干段,把每一段看成一个基向量,这些基向量相应的成分称为子成分.用数学的术语描述就是如下的定义.

定义 2.2 把基向量 ω 和相应的成分向量 x 同样分成 k 段,即有

$$x = \begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(k)} \end{pmatrix} \begin{matrix} n_1 \\ \vdots \\ n_k \end{matrix}, \quad \omega = \begin{pmatrix} \omega_{(1)} \\ \vdots \\ \omega_{(k)} \end{pmatrix} \begin{matrix} n_1 \\ \vdots \\ n_k \end{matrix}.$$

$$n_1 + n_2 + \cdots + n_k = n + 1.$$

令 $S_{(i)} = x_{(i)}(\mathbb{1}'x_{(i)})^{-1}, i=1,2,\cdots,k$,则称 $S_{(i)}$是 x 的子成分.

当 $n_i>1$ 时,$S_{(i)}$是一个向量;当 $n_i=1$ 时,$S_{(i)}=1$.这表明成分向量 x 的子成分在 $n_i=1$ 时需要特殊考虑,这一点在今后的讨论中会经常遇到.另一方面,从子成分的定义还可以看出子成分 $S_{(i)}$有以下的性质:

(1)$S_{(i)} = \omega_{(i)}(\mathbb{1}'\omega_{(i)})^{-1}$.

这是因为 $S_{(i)} = x_{(i)}(\mathbb{1}'x_{(i)})^{-1} = tx_{(i)}(t\,\mathbb{1}'x_{(i)})^{-1}$,而 $tx_{(i)} = w_{(i)}$,于是就有 $S_{(i)} = \omega_{(i)}(\mathbb{1}'\omega_{(i)})^{-1}$.这告诉我们子成分 $S_{(i)}$既是 $x_{(i)}$的成分也是 $\omega_{(i)}$的成分,$\omega_{(i)}$的有些性质可以在 $S_{(i)}$上反映出来.

(2)$S_{(i)}$的对数对比一定是 x 的对数对比,也是 $\omega_{(i)}$的对数对比.

这是由于对数对比的两个性质推出的,这两个性质是:成分 x 的对数对比一定是基 ω 的同一个对数对比(同一个指系数向量相

同)；ω 的一部分 $\omega_{(i)}$ 的对数对比也是 ω 的一个对数对比.用数学的式子表示就是：

因为当 $a'\mathbb{1}=0$ 时，

$$a'\ln x = a'\ln(\omega/t) = a'\ln\omega - a'\mathbb{1}\ln t$$
$$= a'\ln\omega;$$

当 $a'_{(i)}\mathbb{1}_{n_i}=0$ 时，$a'_{(i)}\ln\omega_{(i)}=(0,0,\cdots,a'_{(i)},0,\cdots,0)\ln\omega$，取 $a=(0,0,\cdots,a'_{(i)},0,\cdots,0)$，则有

$$a'\mathbb{1}_{n+1} = a'_{(i)}\mathbb{1}_{n_i} = 0, a'_{(i)}\ln\omega_{(i)} = a'\ln\omega.$$

一个对比的系数是一个 $n+1$ 维的向量，它满足齐次方程 $a'\mathbb{1}=0$，也即 $\mathbb{1}'a=0$. 这一齐次方程的全部解是 $R(\mathbb{1})$ 的正交补空间 $R^{\perp}(\mathbb{1})$，是一个 n 维的子空间，它的一组基是 $\begin{pmatrix} I_n \\ -\mathbb{1}' \end{pmatrix}$ 这一矩阵的 n 个列向量，它相应的投影矩阵是

$$I_{n+1} - P_{\mathbb{1}} = I_{n+1} - \frac{1}{n+1}\mathbb{1}\mathbb{1}'.$$

容易验证：

$$\left(I_{n+1} - \frac{1}{n+1}\mathbb{1}\mathbb{1}'\right)\mathbb{1} = 0,$$

因此任一 $a=\left(I_{n+1}-\frac{1}{n+1}\mathbb{1}\mathbb{1}'\right)x, x\in R_{n+1}$，就一定有性质 $a'\mathbb{1}=0$，也即 a 是一个对比系数；反之，任一对比系数 a，它一定可以表成 $\left(I_{n+1}-\frac{1}{n+1}\mathbb{1}\mathbb{1}'\right)b$ 或 $\begin{pmatrix} I_n \\ -\mathbb{1}' \end{pmatrix}g$. 这些在今后的讨论中都会用到.

§3. 多元正态分布

这一节概要地叙述多元正态分布的一些基本的性质，一些较长的证明都不列出，读者可以参阅一般多元统计分析的教材.

n 维多元标准正态分布用 $N(0, I_n)$ 表示，它的密度函数写成

$\varphi(x)$，$\varphi(x)$的表达式是

$$\varphi(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}x'x}. \tag{3.1}$$

容易看出，当随机向量 $\underset{n\times 1}{x}$ 遵从 $N(0,I_n)$分布时，它的分量 $x_1,\cdots,x_n$ 是独立、同分布 $N(0,1)$的随机变量，期望值为 0，方差是 1. $N(0,1)$是一元的标准正态分布.

我们用 Ex 表示随机向量 x 的期望值组成的向量，$\mathrm{Var}(x)$表示 x 相应协方差矩阵，即

$$Ex = \begin{pmatrix} Ex_1 \\ \vdots \\ Ex_n \end{pmatrix},$$

$$\mathrm{Var}(x) = E(x - Ex)(x - Ex)'.$$

当 $x \sim N(0,I_n)$时，若 $y = \underset{m\times n}{A}x + \mu$，则有

$$\begin{cases} Ey = AEx + \mu = \mu, \\ \mathrm{Var}(y) = AE(x - Ex)(x - Ex)'A' = AA'. \end{cases} \tag{3.2}$$

只要 $AA'>0$，y 就有密度，记$\Sigma = AA'$后，y 的分布密度是

$$\left(\frac{1}{\sqrt{2\pi}}\right)^m \left|\Sigma\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)}. \tag{3.3}$$

这是一般 m 维多元正态分布，记为 $N(\mu,\Sigma)$，μ 的维数或Σ的阶数就是分布的维数. 对于 $N(\mu,\Sigma)$我们用引理的形式列举它的性质，并给以扼要的证明.

引理 3.1 设 $\underset{m\times 1}{y} \sim N(\mu,\Sigma)$，$\Sigma>0$，则有

(1)$\Sigma^{-\frac{1}{2}}(y-\mu) \sim N(0,I_m)$；

(2)$(y-\mu)'\Sigma^{-1}(y-\mu) \sim \chi^2(m)$.

证明 当$\Sigma>0$时，它的特征根 $\lambda_1,\cdots,\lambda_m$ 均为正数，λ_i^α 对任一实数α 均有意义. 由对角化的定理知道存在正交阵 Γ 使

$$\Sigma = \Gamma \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix} \Gamma'.$$

我们规定

$$\sum^{a}=\Gamma\begin{pmatrix}\lambda_1^{a} & & 0\\ & \ddots & \\ 0 & & \lambda_m^{a}\end{pmatrix}\Gamma',$$

于是$\sum^{\frac{1}{2}}$,$\sum^{-\frac{1}{2}}$均有意义,且$\sum^{\frac{1}{2}}\sum^{-\frac{1}{2}}=\sum^{-\frac{1}{2}}\sum^{\frac{1}{2}}=I_m$.

由于 y 的线性函数依然是标准正态分布随机变量 x 的线性函数,因此它还是正态分布,只需求出它的期望值与协方差矩阵,就可以完全确定它的分布,而

$$E\sum^{-\frac{1}{2}}(y-\mu)=\sum^{-\frac{1}{2}}(E(y)-\mu)=0,$$

$$\mathrm{Var}\Big(\sum^{-\frac{1}{2}}(y-\mu)\Big)=\sum^{-\frac{1}{2}}\mathrm{Var}(y)\sum^{-\frac{1}{2}}=I_m,$$

这就证明了(1).由 $x=\sum^{-\frac{1}{2}}(y-\mu)\sim N(0,I_m)$,知道

$$(y-\mu)'\sum^{-1}(y-\mu)=x'x=\sum_{i=1}^{m}x_i^2,$$

因此从 $x_1,\cdots,x_m$ 独立同分布 $N(0,1)$得到

$$x'x\sim\chi^2(m).$$

这就证明了(2).

引理 3.2 若 $y\sim N(\mu,\sum)$,将 y,μ 与$\sum$作相应的分块,即

$$y=\begin{pmatrix}y_{(1)}\\ y_{(2)}\end{pmatrix},\quad \mu=\begin{pmatrix}\mu_{(1)}\\ \mu_{(2)}\end{pmatrix},\quad \sum=\begin{pmatrix}\sum_{11} & \sum_{12}\\ \sum_{21} & \sum_{22}\end{pmatrix},$$

则有

(1)$y_{(i)}\sim N(\mu_{(i)},\sum_{ii}),i=1,2$;

(2)$y_{(1)}$对 $y_{(2)}$的条件分布还是正态,且

$$\begin{cases}E\{y_{(1)}\mid y_{(2)}\}=\mu_{(1)}+\sum_{12}\sum_{22}^{-1}(y_{(2)}-\mu_{(2)}),\\ \mathrm{Var}(y_{(1)}\mid y_{(2)})=\sum_{11}-\sum_{12}\sum_{22}^{-1}\sum_{21}.\end{cases}\tag{3.4}$$

证明 这只需将(3.3)的分布密度用边缘密度和条件分布密度乘积的形式写出就得.关键是指数上二次型的分解,用到了一个

矩阵分块求逆的公式，就是

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \Sigma_{11}^{-1}\Sigma_{12} \\ -I \end{pmatrix} \times \Sigma_{22\cdot 1}^{-1}\left(\Sigma_{21}\Sigma_{11}^{-1} \quad -I\right),$$

其中

$$\Sigma_{22\cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

用这个公式，就得

$$\begin{aligned}
&(y-\mu)'\Sigma^{-1}(y-\mu) \\
&= ((y_{(1)}-\mu_{(1)})'(y_{(2)}-\mu_{(2)})') \\
&\quad \times \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} y_{(1)}-\mu_{(1)} \\ y_{(2)}-\mu_{(2)} \end{pmatrix} \\
&= (y_{(1)}-\mu_{(1)})'\Sigma_{11}^{-1}(y_{(1)}-\mu_{(1)}) \\
&\quad + \left(y_{(2)}-\mu_{(2)}-\Sigma_{21}\Sigma_{11}^{-1}(y_{(1)}-\mu_{(1)})\right) \\
&\quad \times \Sigma_{22\cdot 1}^{-1}\Big(y_{(2)}-\mu_{(2)} \\
&\quad - \Sigma_{21}\Sigma_{11}^{-1}(y_{(1)}-\mu_{(1)})\Big) \\
&\triangleq Q_1 + Q_2.
\end{aligned}$$

于是分布密度(3.3)可写成

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \left(|\Sigma_{11}|\,|\Sigma_{22\cdot 1}|\right)^{-\frac{1}{2}} e^{-\frac{1}{2}Q_1} e^{-\frac{1}{2}Q_2},$$

其中

$$\begin{aligned}
Q_1 &= (y_{(1)}-\mu_{(1)})'\Sigma_{11}^{-1}(y_{(1)}-\mu_{(1)}), \\
Q_2 &= (y_{(2)}-a)'\Sigma_{22\cdot 1}^{-1}(y_{(2)}-a), \\
a &= \mu_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}(y_{(1)}-\mu_{(1)}),
\end{aligned}$$

就能得所要的结论.$\sum$分块后,相应的有

$$\left|\sum\right| = \begin{vmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{vmatrix}$$
$$= \left|\sum_{11}\right| \left|\sum_{22} - \sum_{21}\sum_{11}^{-1}\sum_{12}\right|$$
$$= \left|\sum_{11}\right| \left|\sum_{22\cdot 1}\right|,$$

同样地有$\left|\sum\right| = \left|\sum_{22}\right| \left|\sum_{11} - \sum_{12}\sum_{22}^{-1}\sum_{21}\right| = \left|\sum_{22}\right| \left|\sum_{11\cdot 2}\right|$. 对分块求逆公式也是类似地还有另一表达式.从密度(3.3)分块后写出的乘积就可得:

$$y_{(1)} \sim N\left(\mu_{(1)}, \sum_{11}\right),$$

$$y_{(2)} \mid y_{(1)} \sim N\left(\mu_{(2)} + \sum_{21}\sum_{11}^{-1}(y_{(1)} - \mu_{(1)}), \sum_{22\cdot 1}\right),$$

将脚标1与2对换,就可得$y_{(2)}$与$y_{(1)} \mid y_{(2)}$的分布,于是就证明了本引理.

现在来讨论一个特殊的分布,它与第二章中单形上的逻辑正态分布有密切的关系.

假定$y \sim N\left(0, \sum\limits_{m \times m}\right), \sum > 0$,这里假定期望为0是避免推导过程中形式上过于庞杂.考虑y的$m-1$个对比,求这些对比的联合分布密度.

令$F = (I_{m-1} - \mathbb{1})$, $z = Fy$. 注意此时z是$m-1$维的向量,实际上

$$z_i = y_i - y_m, i = 1, 2, \cdots, m-1.$$

现在的问题是要求z的分布密度.由于z是y的线性函数,因而z也是正态分布.要写出z的密度,只需求z的期望Ez, z的协差阵$\mathrm{Var}(z)$,然后再求得$\mathrm{Var}(z)$的逆就行了.今

$$Ez = FEy = 0,$$
$$\mathrm{Var}(z) = F\mathrm{Var}(y)F' = F\sum F'.$$

如何求出$F\sum F'$的逆阵就成了求z分布的核心问题.解决这个问

题需要再一次应用分块求逆的公式. 当我们已知 $y \sim N(0, \Sigma)$ 的密度时,实际上我们已知的是 Σ^{-1} 的表达式,如将 Σ^{-1} 和 Σ 一样分块写出,记为

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \Sigma^{-1} = \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix},$$

则由分块求逆公式可知

$$\Sigma^{22} = \Sigma_{22\cdot 1}^{-1} = \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1},$$

也即有

$$\left(\Sigma^{22}\right)^{-1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12},$$

或

$$\left(\Sigma^{11}\right)^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

如将 Σ^{-1} 作为已知的,要求 Σ 中的 Σ_{ii},则有

$$\Sigma_{11}^{-1} = \Sigma^{11} - \Sigma^{12}\left(\Sigma^{22}\right)^{-1}\Sigma^{21}. \tag{3.5}$$

现在我们用(3.5)来求 $(F\Sigma F')^{-1}$. 考虑

$$\begin{pmatrix} I_m & -\mathbb{1} \\ \mathbb{1}' & 1 \end{pmatrix} \Sigma \begin{pmatrix} I_m & -\mathbb{1} \\ \mathbb{1}' & 1 \end{pmatrix}' = \begin{pmatrix} F \\ \mathbb{1}' \end{pmatrix} \Sigma (F' \ \mathbb{1})$$

$$= \begin{pmatrix} F\Sigma F' & F\Sigma \mathbb{1} \\ \mathbb{1}'\Sigma F' & \mathbb{1}'\Sigma \mathbb{1} \end{pmatrix} \triangleq A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

很明显,我们要求的 $(F\Sigma F')^{-1}$ 正好是(3.5)中相应 Σ_{11}^{-1}. 因此用公式(3.5),就得

$$\left(F\Sigma F'\right)^{-1} = A_{11}^{-1} = A^{11} - A^{12}(A^{22})^{-1}A^{21},$$

由于

$$\begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix} = A^{-1} = (F' \ \mathbb{1})^{-1}\Sigma^{-1}\begin{pmatrix} F \\ \mathbb{1}' \end{pmatrix}^{-1},$$

而 $\begin{pmatrix} F \\ \mathbb{1}' \end{pmatrix}^{-1}$ 仍可以用分块求逆公式得到,因为

$$\begin{pmatrix} F \\ \mathbb{1}' \end{pmatrix}^{-1} = \begin{pmatrix} I_{m-1} & -\mathbb{1} \\ \mathbb{1}' & 1 \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} I_{m-1} & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{m}\begin{pmatrix} -\mathbb{1} \\ -1 \end{pmatrix}(\mathbb{1}' \ -1)$$

$$= \begin{pmatrix} I_{m-1} - \dfrac{1}{m}\mathbb{1}\mathbb{1}' & \dfrac{1}{m}\mathbb{1} \\ -\dfrac{1}{m}\mathbb{1}' & \dfrac{1}{m} \end{pmatrix}.$$

于是

$$\begin{pmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{pmatrix} = \begin{pmatrix} I_{m-1} - \dfrac{1}{m}\mathbb{1}\mathbb{1}' & -\dfrac{1}{m}\mathbb{1} \\ \dfrac{1}{m}\mathbb{1}' & \dfrac{1}{m} \end{pmatrix} \begin{pmatrix} \sum^{11} & \sum^{12} \\ \sum^{21} & \sum^{22} \end{pmatrix}$$

$$\times \begin{pmatrix} I_{m-1} - \dfrac{1}{m}\mathbb{1}\mathbb{1}' & \dfrac{1}{m}\mathbb{1} \\ -\dfrac{1}{m}\mathbb{1}' & \dfrac{1}{m} \end{pmatrix}.$$

因此得到

$$A^{11} = \left(I_{m-1} - \frac{1}{m}\mathbb{1}\mathbb{1}' \quad -\frac{1}{m}\mathbb{1}\right)\begin{pmatrix} \sum^{11} & \sum^{12} \\ \sum^{21} & \sum^{22} \end{pmatrix}\begin{pmatrix} I_{m-1} - \dfrac{1}{m}\mathbb{1}\mathbb{1}' \\ -\dfrac{1}{m}\mathbb{1}' \end{pmatrix}$$

$$= \left[(I_{m-1} \quad 0) - \frac{1}{m}\mathbb{1}\mathbb{1}'\right]\sum^{-1}\left[\begin{pmatrix} I_{m-1} \\ 0 \end{pmatrix} - \frac{1}{m}\mathbb{1}\mathbb{1}'\right],$$

$$A^{12} = \left[(I_{m-1} \quad 0) - \frac{1}{m}\mathbb{1}\mathbb{1}'\right]\sum^{-1}\left(\frac{1}{m}\mathbb{1}\right) = (A^{21})',$$

$$A^{22} = \frac{1}{m^2}\mathbb{1}'\sum^{-1}\mathbb{1}.$$

而

$$\left(F\sum F'\right)^{-1} = A_{11}^{-1} = A^{11} - A^{12}(A^{22})^{-1}A^{21}$$

$$= \left[(I_{m-1} \quad 0) - \frac{1}{m}\mathbb{1}\mathbb{1}'\right]\sum^{-1}\left[\begin{pmatrix} I_{m-1} \\ 0 \end{pmatrix} - \frac{1}{m}\mathbb{1}\mathbb{1}'\right]$$

$$
\begin{aligned}
&-\left[(I_{m-1}\quad 0)-\frac{1}{m}\mathbb{1}\mathbb{1}'\right]\Sigma^{-1}\frac{1}{m}\mathbb{1}\frac{m^2}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}\\
&\times\frac{1}{m}\mathbb{1}'\Sigma^{-1}\left[\begin{pmatrix}I_{m-1}\\0\end{pmatrix}-\frac{1}{m}\mathbb{1}\mathbb{1}'\right]\\
=&\left[(I_{m-1}\quad 0)-\frac{1}{m}\mathbb{1}\mathbb{1}'\right]\Sigma^{-1}\left[\begin{pmatrix}I_{m-1}\\0\end{pmatrix}-\frac{1}{m}\mathbb{1}\mathbb{1}'\right]\\
&-\left[(I_{m-1}\quad 0)-\frac{1}{m}\mathbb{1}\mathbb{1}'\right]\frac{\Sigma^{-1}\mathbb{1}\mathbb{1}'\Sigma^{-1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}\\
&\times\left[\begin{pmatrix}I_{m-1}\\0\end{pmatrix}-\frac{1}{m}\mathbb{1}\mathbb{1}'\right]\\
=&(I_{m-1}\quad 0)\left[\Sigma^{-1}-\frac{\Sigma^{-1}\mathbb{1}\mathbb{1}'\Sigma^{-1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}\right]\begin{pmatrix}I_{m-1}\\0\end{pmatrix}.
\end{aligned}
$$

令

$$
B=\Sigma^{-1}-\frac{\Sigma^{-1}\mathbb{1}\mathbb{1}'\Sigma^{-1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}\triangleq\begin{array}{cc}\begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix} & \begin{matrix}m-1\\1\end{matrix}\\ \begin{matrix}m-1 & \quad 1\end{matrix} & \end{array},
$$

则就得：

若 $y\sim N(0,\Sigma)$，$z=Fy$，$F=(I_{m-1}\quad -\mathbb{1})$，

于是 $z\sim N(0,F\Sigma F')$，z 的密度为

$$
\left(\frac{1}{\sqrt{2\pi}}\right)^{m-1}|B_{11}|^{\frac{1}{2}}e^{-\frac{1}{2}z'B_{11}z}. \tag{3.6}
$$

容易验证，矩阵 B 是退化的，它的秩是 $m-1$，而且

$$
B\mathbb{1}=\Sigma^{-1}\mathbb{1}-\frac{\Sigma^{-1}\mathbb{1}\mathbb{1}'\Sigma^{-1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}\mathbb{1}=0.
$$

有关的进一步讨论我们留作习题. 上述结果归结为

定理 3.1 若 $y\sim N\left(0,\underset{m\times m}{\Sigma}\right)$，$z=Fy$，$F=(I_{m-1}\quad -\mathbb{1})$，则令 $B=\Sigma^{-1}-\frac{\Sigma^{-1}\mathbb{1}\mathbb{1}'\Sigma^{-1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}=\begin{array}{cc}\begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix} & \begin{matrix}m-1\\1\end{matrix}\\ \begin{matrix}m-1 & \quad 1\end{matrix} & \end{array}$ 后，就有

$$z \sim N(0, B_{11}^{-1}).$$

与正态分布有关的另一个分布是反正态分布(inverse gaussian distribution).不要从名词上误解为它是服从正态分布的随机变量 ξ 的逆 ξ^{-1} 的分布,取这个名词是因为它的特征函数恰好是正态的反函数,这一点在下面求得特征函数就可以说明了.尽管反正态分布在 50 年代中期就提出来了,但并未引起重视,直到 70 年代,一些实际应用表明这一分布是很有意义的,进一步的研究表明它与过程的一些特征量有关.我们国内这一方面在教材和论文中还没有出现过这一类分布,因此本书稍多地给以一些介绍,以期引起理论和应用方面的兴趣,正文中我们只论述与成分数据分析的有关内容,在附录(见本章习题后的附录)中,对它的进一步推广和应用作一些介绍.

在处理反正态分布时,需要一种求积分的微商方法,我们先引入一个积分求值的公式,以它作为工具,导出反正态分布的密度以及有关的统计性质.

首先我们证明

$$\int_0^\infty t^{-\frac{1}{2}} e^{-\left(t+\frac{a}{t}\right)} dt = \sqrt{\pi} e^{-2\sqrt{a}}, \tag{3.7}$$

$\forall\ a \geqslant 0$ 都成立.

显然,$a=0$ 时,(3.7)式是 Γ 函数的一个熟知的等式,因 $a=0$ 时上式是成立的.记(3.7)式左端为 $F(a)$,它是 a 的函数,易见有:对 $a>0$

$$\frac{dF(a)}{da} = \int_0^\infty t^{-\frac{1}{2}}(-t)^{-1} e^{-\left(t+\frac{a}{t}\right)} dt.$$

令 $x=\frac{a}{t}$,则 $t=\frac{a}{x}$ 且 $dt=-\frac{a}{x^2}dx$,代入上式右端,得到

$$\begin{aligned}\frac{dF(a)}{da} &= -\frac{1}{\sqrt{a}}\int_0^\infty x^{-\frac{1}{2}} e^{-\left(x+\frac{a}{x}\right)} dx \\ &= -\frac{1}{\sqrt{a}} F(a).\end{aligned}$$

注意到 $F(0)=\sqrt{\pi}$,于是解上述微分方程就得(3.7)式.

将(3.7)式两端均对 a 求微商,就得

$$\int_0^{\infty} t^{-\frac{3}{2}} e^{-\left(t+\frac{a}{t}\right)} dt = \sqrt{\frac{\pi}{a}} e^{-2\sqrt{a}}. \tag{3.8}$$

实际上,(3.8)式也可以通过积分变量的变换得到,只要留心上面导出微分方程时的计算,就可以马上知道相应的变换是什么.从(3.7),(3.8)式可以导出另两个等式,我们只证其中之一,另一个留给读者.今对 $b>0$ 有

$$\begin{aligned}
&\int_0^{\infty} t^{-\frac{1}{2}} e^{-b\left(t+\frac{a}{t}\right)} dt \\
&= \int_0^{\infty} \sqrt{b}(bt)^{-\frac{1}{2}} e^{-\left(bt+\frac{ab^2}{bt}\right)} \frac{1}{b} d(bt) \\
&= \frac{1}{\sqrt{b}} \int_0^{\infty} u^{-\frac{1}{2}} e^{-\left(u+\frac{ab^2}{u}\right)} du \\
&= \sqrt{\frac{\pi}{b}} e^{-2b\sqrt{a}}.
\end{aligned}$$

因此有

$$\begin{cases}
\int_0^{\infty} t^{-\frac{1}{2}} e^{-b\left(t+\frac{a}{t}\right)} dt = \sqrt{\frac{\pi}{b}} e^{-2b\sqrt{a}}, \\
\int_0^{\infty} t^{-\frac{3}{2}} e^{-b\left(t+\frac{a}{t}\right)} dt = \sqrt{\frac{\pi}{ab}} e^{-2b\sqrt{a}}.
\end{cases} \tag{3.9}$$

注意到

$$\begin{aligned}
b\left(t+\frac{a}{t}\right) &= b\left(t-2\sqrt{a}+\frac{a}{t}\right)+2b\sqrt{a} \\
&= \frac{b}{t}(t-\sqrt{a})^2+2b\sqrt{a},
\end{aligned}$$

从(3.9)式就得到

$$\begin{cases}
\sqrt{\frac{b}{\pi}} \int_0^{\infty} t^{-\frac{1}{2}} e^{-\frac{b}{t}(t-\sqrt{a})^2} dt = 1, \\
\sqrt{\frac{ab}{\pi}} \int_0^{\infty} t^{-\frac{3}{2}} e^{-\frac{b}{t}(t-\sqrt{a})^2} dt = 1.
\end{cases} \tag{3.10}$$

从(3.10)就可以引出反正态分布.若随机变量 ξ 的分布密度

$$p(t)=\sqrt{\frac{\lambda}{2\pi}}t^{-\frac{3}{2}}\mathrm{e}^{-\frac{\lambda}{2\mu^2 t}(t-\mu)^2},t>0, \qquad (3.11)$$

其中参数 $\mu>0,\lambda>0$,则称 ξ 遵从反正态分布,用 $\mathrm{IN}(\mu,\mu^3/\lambda)$ 表示. 下面我们将证明 μ 是 ξ 的期望值,μ^3/λ 是 ξ 的方差,因此这一记号与正态的写法一致. 在求期望方差之前,我们先证一条引理.

引理 3.3 若 $\xi\sim\mathrm{IN}(\mu,\mu^3/\lambda)$,则它的逆 ξ^{-1} 的分布密度为

$$\sqrt{\frac{\lambda}{2\pi}}t^{-\frac{1}{2}}\mathrm{e}^{-\frac{\lambda}{2t}\left(t-\frac{1}{\mu}\right)^2},\quad t>0. \qquad (3.12)$$

证明 注意(3.10)中第二个等式中取 $a=\mu^2,b=\frac{\lambda}{2\mu^2}$,左端被积函数就是 $\mathrm{IN}(\mu,\mu^3/\lambda)$ 的密度. 而 $P(\xi^{-1}<t)=P(\xi>t^{-1})$ $=\int_{t^{-1}}^{\infty}\sqrt{\frac{\lambda}{2\pi}}x^{-\frac{3}{2}}\mathrm{e}^{-\frac{\lambda}{2x\mu^2}(x-\mu)^2}\mathrm{d}x$,两边对 t 求微商,就得所要的结论.

(3.12)相应分布称为逆反正态分布,意思是它是反正态分布 $\mathrm{IN}(\mu,\mu^3/\lambda)$ 随机变量 ξ 之逆 ξ^{-1} 的分布. 因为它的英文名称是 reciprocal inverse gaussian distribution,它的记号记为 $\mathrm{RIN}(\mu,\mu^3/\lambda)$. 很明显,$E\xi^{-1}$ 并不是 μ,方差也不是 μ^3/λ,这一写法重点是表明它是 $\mathrm{IN}(\mu,\mu^3/\lambda)$ 分布随机变量之逆的分布. reciprocal 一词还有相互的意义,实际上也是,当 $\eta\sim\mathrm{RIN}(\mu,\mu^3/\lambda)$ 时,η^{-1} 就以反正态分布 $\mathrm{IN}(\mu,\mu^3/\lambda)$ 为分布,它们是相互为逆分布,因此今后有关的结论我们都成对给出. 我们往往只对一个分布给出证明,另一个分布因为方法类似就不证了,读者自己去证.

比较(3.10)和(3.11),(3.12),可以看出反正态分布的参数可以有不同的选择,(3.10)是用 a,b 两个参数,(3.11),(3.12)是用 λ,μ 这两个参数,它们之间的关系是

$$a=\mu^2,b=\frac{\lambda}{2\mu^2}=\frac{\lambda}{2a}.(\text{即 }2ab=\lambda)$$

有时为了证明简洁,选用 a,b 或 λ,μ 都是可以的,λ,μ 的统计意义比较明显. 下面我们来求分布的矩母函数,从而导出它们的期望、方差的表达式以及其他的性质.

定理 3.2 若 $\xi \sim \mathrm{IN}(\mu, \mu^3/\lambda)$,则有

$$\begin{cases} g_\xi(\theta) = Ee^{\xi\theta} = \exp\left\{\dfrac{\lambda}{\mu}\left(1 - \sqrt{1 - 2\mu^2\theta/\lambda}\right)\right\}, \\ g_{\xi^{-1}}(\theta) = Ee^{\xi^{-1}\theta} = \left(1 - \dfrac{2\theta}{\lambda}\right)^{-\frac{1}{2}}\exp\left\{\dfrac{\lambda}{\mu}\left(1 - \sqrt{1 - 2\theta/\lambda}\right)\right\}. \end{cases} \tag{3.13}$$

证明 令

$$\begin{aligned} g_\xi(\theta) &= Ee^{\xi\theta} \\ &= \int_0^\infty t^{-\frac{3}{2}}\sqrt{\frac{ab}{\pi}}e^{2b\sqrt{a}+\theta t-b\left(t+\frac{a}{t}\right)}\mathrm{d}t \\ &= \sqrt{\frac{ab}{\pi}}e^{2b\sqrt{a}}\int_0^\infty t^{-\frac{3}{2}}e^{-(b-\theta)\left(t+\frac{ab}{b-\theta}t^{-1}\right)}\mathrm{d}t \\ &= \sqrt{\frac{ab}{\pi}}e^{2b\sqrt{a}}\cdot\sqrt{\frac{\pi}{ab}}\cdot e^{-2\sqrt{ab(b-\theta)}} \\ &= e^{2b\sqrt{a}\left(1-\sqrt{1-\theta/b}\right)}. \end{aligned}$$

用 $a = \mu^2, b = \dfrac{\lambda}{2\mu^2}$代入上式右端,就是(3.13)第一式.类似方法可以证得第二式.

现在来说明反正态分布的命名.正态分布 $N(\mu, \sigma^2)$的矩母函数是 $e^{\mu t + \frac{\sigma^2}{2}t^2}$,它取对数后就是 $\mu t + \dfrac{\sigma^2}{2}t^2$,令它为 $-\theta$,就得

$$\frac{\sigma^2}{2}t^2 + \mu t + \theta = 0.$$

将 θ 看作常数,t 看作变元,解上述一元二次方程,得到

$$t = \frac{-\mu \pm \sqrt{\mu^2 - 2\sigma^2\theta}}{\sigma^2} = -\frac{\mu}{\sigma^2}\left(1 \pm \sqrt{1 - 2\sigma^2\theta/\mu^2}\right).$$

很明显,上式右端与(3.13)中 $g_\xi(\theta)$的对数完全相似,在这个意义上,它与正态的正好相反.

有了定理 3.2,可以得到一系列重要的推论,我们逐一给以说明.

系 1 若 $\xi \sim \mathrm{IN}(\mu, \mu^3/\lambda)$,则有

$$\begin{cases} E(\xi) = \mu, \mathrm{Var}(\xi) = \mu^3/\lambda, \\ E(\xi^{-1}) = \dfrac{1}{\mu} + \dfrac{1}{\lambda}; \mathrm{Var}(\xi) = \dfrac{\lambda + 2\mu}{\lambda^2 \mu}. \end{cases} \tag{3.14}$$

利用 $\ln g_\xi(\theta)$ 展开式的 θ 与 θ^2 的系数，就可求出 ξ 的期望与方差. 因为

$$\begin{aligned} \ln g_\xi(\theta) &= \frac{\lambda}{\mu}(1 - (1 - 2\mu^2\theta/\lambda)^{\frac{1}{2}}) \\ &= \frac{\lambda}{\mu}\left(1 - 1 + \frac{1}{2}\frac{2\mu^2\theta}{\lambda} + \frac{1}{2}\cdot\frac{1}{2}\left(\frac{1}{2} - 1\right)\left(\frac{2\mu^2\theta}{\lambda}\right)^2 + \cdots\right) \\ &= \mu\theta + \frac{1}{2}\frac{\mu^3}{\lambda}\theta^2 + \cdots \end{aligned}$$

同样的方法对 $\ln g_{\xi^{-1}}(\theta)$ 展开就得 $E(\xi^{-1})$ 与 $\mathrm{Var}(\xi^{-1})$.

系 2 若 $x_1, \cdots, x_n$ 独立同分布 $\mathrm{IN}(\mu, \mu^3/\lambda)$，则 $S = x_1 + \cdots + x_n \sim \mathrm{IN}(n\mu, n\mu^3/\lambda)$.

证明
$$\begin{aligned} Ee^{\theta S} &= Ee^{\theta(x_1 + \cdots + x_n)} = \prod_{i=1}^{n} Ee^{\theta x_i} = (Ee^{\theta x_1})^n \\ &= e^{n\frac{\lambda}{\mu}(1 - \sqrt{1 - 2\mu^2\theta/\lambda})}. \end{aligned}$$

由于 $ES = nEx_1 = n\mu$, $\mathrm{Var}(S) = n\mathrm{Var}(x_1) = n\mu^3/\lambda$，注意到 $\mathrm{IN}(n\mu, n\mu^3/\lambda)$ 相应的矩母函数是在表达式中用 $n\mu$ 代 μ，$n^2\lambda$ 代 λ 就可由 $g_\xi(\theta)$ 得到，即为

$$\begin{aligned} &\exp\left\{\frac{n^2\lambda}{n\mu}\left(1 - \sqrt{1 - 2(n\mu)^2\theta/n^2\lambda}\right)\right\} \\ &= \exp\left\{n\frac{\lambda}{\mu}\left(1 - \sqrt{1 - 2\mu^2\theta/\lambda}\right)\right\}. \end{aligned}$$

它就是 $Ee^{\theta S}$，这样就证明了系 2.

从证明中可以看出，对 $\mathrm{IN}(\mu, \mu^3/\lambda)$，如果以 (μ, β) 为参数，$\beta = \dfrac{\lambda}{\mu^2} = \dfrac{E\xi}{\mathrm{Var}(\xi)}$，于是 $\mathrm{IN}(\mu, \mu^3/\lambda)$ 可以写成 $\mathrm{IN}(\mu, \mu/\beta)$，相应的矩母函数写成

$$g_\xi(\theta) = \exp\{\mu\beta(1 - \sqrt{1 - 2\theta/\beta})\},$$

于是有

若 $\xi_1,\cdots,\xi_n$ 独立，$\xi_i \sim \mathrm{IN}(\mu_i,\mu_i/\beta)$，$i=1,2,\cdots,n$ 则 $\xi_+=\sum_{i=1}^{n}\xi_i \sim \mathrm{IN}\left(\sum_{i=1}^{n}\mu_i,\sum_{i=1}^{n}\mu_i/\beta\right)$（即 $\mathrm{IN}(\mu_+,\mu_+/\beta)$），其中 $\mu_+=\sum_{i=1}^{n}\mu_i$.

这就告诉我们用(μ,β)作为参数是方便的.

现在来求参数 μ,λ 的估计量. 设 $x_1,\cdots,x_n$ 来自 $\mathrm{IN}(\mu,\mu^3/\lambda)$，这个样本相应的均值和方差是

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

从矩估计法就得（见(3.14)式）

$$\bar{x}=\hat{\mu},s^2=\frac{\hat{\mu}^3}{\hat{\lambda}},$$

因此得

$$\hat{\mu}=\bar{x},\quad \hat{\lambda}=\hat{\mu}^3/s^2=\bar{x}^3/s^2. \tag{3.15}$$

然而从(3.14)第二组等式可以看到

$$E\frac{1}{x_i}=\frac{1}{\mu}+\frac{1}{\lambda},$$

即

$$E\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{-1}\right)=\frac{1}{\mu}+\frac{1}{\lambda}.$$

因此可以导出 λ 的另一个估计量

$$\hat{\lambda}=\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{-1}-\bar{x}^{-1}\right)^{-1},$$

下面我们将看到，它正是最大似然估计.

样本 $x_1,\cdots,x_n$ 的联合密度是 $p(x_1,\cdots,x_n;\mu,\lambda)$，似然函数 $L(\mu,\lambda;x_1,\cdots,x_n)=p(x_1,\cdots,x_n;\mu,\lambda)$，它就是

$$\left(\prod_{i=1}^{n}x_i^{-\frac{3}{2}}\right)\left(\frac{\lambda}{2\pi}\right)^{\frac{n}{2}}\exp\left\{-\frac{\lambda}{2\mu^2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{x_i}\right\}.$$

于是

$$\ln L(\mu,\lambda;x_1,\cdots,x_n) = -\frac{n}{2}\ln 2\pi + \frac{n}{2}\ln\lambda - \frac{3}{2}\sum_{i=1}^{n}\ln x_i - \frac{\lambda}{2\mu^2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{x_i}$$

对上式两端分别微商,就导出似然方程

$$0 = \frac{\partial \ln L(\mu,\lambda;x_1,\cdots,x_n)}{\partial x} = \frac{n}{2\lambda} - \frac{1}{2\mu^2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{x_i}$$

$$0 = \frac{\partial \ln L(\mu,\lambda;x_1,\cdots,x_n)}{\partial \mu} = \frac{\lambda}{\mu^3}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{x_i} + \frac{\lambda}{\mu^2}\sum_{i=1}^{n}\frac{x_i-\mu}{x_i}$$

$$= \frac{\lambda}{\mu^3}\left(n\bar{x} - 2n\mu + \mu^2\sum_{i=1}^{n}\frac{1}{x_i} + n\mu - \mu^2\sum_{i=1}^{n}\frac{1}{x_i}\right)$$

$$= \frac{\lambda}{\mu^3}(n(\bar{x}-\mu)).$$

因此解得最大似然估计为

$$\hat{\mu} = \bar{x}, \quad \hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^{-1} - \bar{x}^{-1}\right)^{-1}. \qquad (3.16)$$

如果用 h_n 表示 $x_1,\cdots,x_n$ 的调和平均,即

$$h_n = \left(\frac{1}{n}\sum x_i^{-1}\right)^{-1},$$

或

$$h_n^{-1} = \frac{1}{n}\sum_{i=1}^{n}x_i^{-1}$$

由算术平均、调和平均不等式知道

$$h_n \leqslant \bar{x},$$

也即有

$$h_n^{-1} \geqslant \bar{x}^{-1}.$$

而 $\hat{\lambda} = (h_n^{-1} - \bar{x}^{-1})^{-1}$,它具有 $\hat{\lambda} \geqslant 0$ 的性质,所以是一个合乎非负要求的估计.

§4. 对数正态分布

随机变量 $\omega>0$,如果它的对数 $\ln\omega$ 遵从正态分布,则称 ω 遵

从对数正态分布.类似地,如果随机向量 ω 的每一个分量都是正的,各自取对数后形成的随机向量 $\ln\omega$ 遵从多元正态分布,则称 ω 遵从对数正态分布.

为了从多元正态分布的密度导出对数正态的密度,我们先证明一个引理,这一引理在今后有关分布的推导中是经常要用到的.

引理 4.1 假定随机向量 $\underset{n\times 1}{x}$ 的联合密度是已知的 $p(x)$, $p(x)$在区域 D 上不为 0,在 D 以外均为 0.考虑 x 的函数 $y=g(x)$, y 的取值范围是 $\widetilde{D}$,且 $g(x)$是 D 到 $\widetilde{D}$ 上的 1-1 变换, x 可以用 y 表示,即有反函数 $x=h(y)$,当 g,h 可微,而且偏微商均连续时, y 的分布密度 $f(y)$有如下的表达式:

$$f(y) = p(h(y))|J(x|y)|, y \in \widetilde{D}, \tag{4.1}$$

其中 $J(x|y)$是 x 对 y 的雅可比行列式, $|J(x|y)|$表示行列式的绝对值. $f(y)$在 $\widetilde{D}$ 外为 0.

证明 令

$$I(y) = \begin{cases} 1, & y < a, \\ 0, & \text{其他}. \end{cases}$$

即 $I(y)$是集合$\{y: y<a\}$的示性函数, y 与 a 均为 n 维向量.于是

$$\begin{aligned} P(y_1 < a_1, \cdots, y_n < a_n) &= P(y < a) \\ &= P(g(x) < a) \\ &= \underset{\substack{g(x)<a \\ x\in D}}{\int\cdots\int} p(x_1, \cdots, x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n \\ &= \underset{D}{\int\cdots\int} I(g(x))p(x)\mathrm{d}x, \end{aligned}$$

其中 $\mathrm{d}x=\mathrm{d}x_1\cdots\mathrm{d}x_n$.作变数替换,令 $y=g(x)$,则有 $x=h(y)$, $y\in\widetilde{D}$,于是上式右端积分就变成对 y 在 $\widetilde{D}$ 上的积分.即有

$$\begin{aligned} \underset{D}{\int\cdots\int} I(g(x))p(x)\mathrm{d}x &= \underset{\widetilde{D}}{\int\cdots\int} I(y)p(h(y))|J(x|y)|\mathrm{d}y \\ &= \underset{\substack{y<a \\ y\in\widetilde{D}}}{\int\cdots\int} p(h(y))|J(x|y)|\mathrm{d}y_1\cdots\mathrm{d}y_n \end{aligned}$$

这就证明了(4.1)式.

这一引理告诉我们求随机向量函数的分布时,只要函数满足引理的条件,就可以直接写出随机向量函数的分布密度,关键是求出 $h(y)$ 以及 $J(x|y)$.

现在用引理 4.1 来导出对数正态分布的密度函数.

例 4.1 一元对数正态分布.

设 $x \sim N(\mu, \sigma^2)$,考虑 $y = e^x$, 于是 $x = \ln y$. 很明显,此时与引理 4.1 中相应的有

$$D = (-\infty, \infty), \widetilde{D} = (0, \infty), h(y) = \ln y,$$

$$J(x|y) = \frac{d}{dy}\ln y = \frac{1}{y}.$$

因此,由(4.1)得 y 的分布密度为

$$f(y) = \begin{cases} \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\ln y - \mu)^2}, & y > 0, \\ 0, & \text{其他}. \end{cases} \tag{4.2}$$

例 4.2 多元对数正态分布.

设 $\underset{n\times 1}{x} \sim N(\mu, \Sigma), x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$,考虑 $y_i = e^{x_i}, i = 1, 2, \cdots, n$,于是 $x = \ln y$. 此时与引理 4.1 中相应的有

$$D = R_n, \widetilde{D} = R_n^+, h(y) = \ln y,$$

$$J(x \mid y) = \left(\prod_{i=1}^{n} y_i\right)^{-1}.$$

因此,由(4.1)得 y 的分布密度为

$$f(y) = \begin{cases} \left(\frac{1}{\sqrt{2\pi}}\right)^n |\Sigma|^{-\frac{1}{2}} \left(\prod_{i=1}^{n} y_i\right)^{-1} e^{-\frac{1}{2}(\ln y - \mu)' \Sigma^{-1} (\ln y - \mu)}, & y > 0, \\ 0, & \text{其他}. \end{cases} \tag{4.3}$$

由于对数正态分布已有较长的历史,有关的统计性质和应用已有专门的著作,这里列举一些重要的性质并给出简略的证明.

假定 $\underset{n\times 1}{x} \sim N(\mu,\Sigma)$，$x=\ln y$，于是 y 遵从对数正态分布，记为 $y\sim\Lambda(\mu,\Sigma)$，现在来求 y 的期望值、方差协差矩阵等统计参数. 先考虑一元的情形，再考虑多元的情形.

1. $y\sim\Lambda(\mu,\sigma^2)$，则有

$$\begin{cases} Ey = \exp\left(\mu + \dfrac{\sigma^2}{2}\right), \\ \mathrm{Var}(y) = [\exp(2\mu + \sigma^2)][e^{\sigma^2} - 1]. \end{cases} \tag{4.4}$$

证明 当 $x\sim N(\mu,\sigma^2)$时，

$$\begin{aligned} Ee^{tx} &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}[(x-\mu)^2-2t\sigma^2 x]}dx \\ &= e^{\mu t+\frac{1}{2}\sigma^2t^2}\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu-t\sigma^2)^2}dx \\ &= \exp\left\{\mu t + \frac{1}{2}\sigma^2t^2\right\}, \end{aligned}$$

而 $y=e^x$，因此 $Ey^\alpha = Ee^{\alpha x} = \exp\left\{\mu\alpha + \dfrac{1}{2}\sigma^2\alpha^2\right\}$，当 $\alpha=1$ 时，$Ey = \exp\left\{\mu + \dfrac{\sigma^2}{2}\right\}$；当 $\alpha=2$ 时，

$$Ey^2 = \exp\{2\mu + 2\sigma^2\},$$

于是 $\mathrm{Var}(y) = Ey^2 - (Ey)^2 = [\exp\{2\mu+\sigma^2\}][e^{\sigma^2}-1]$.

2. $\underset{n\times 1}{y} \sim \Lambda(\underset{n\times 1}{\mu}, \underset{n\times n}{\Sigma})$，则有

$$\begin{cases} Ey_i = \exp\left\{\mu_i + \dfrac{1}{2}\sigma_{ii}\right\}, i = 1,2,\cdots,n \\ \mathrm{Var}(y_i) = [\exp\{2\mu_i + \sigma_{ii}\}][e^{\sigma_{ii}} - 1], i = 1,2,\cdots,n \end{cases} \tag{4.5}$$

这从(4.4)可以直接导出. 令 $x=\ln y\sim N(\mu,\Sigma)$，且

$$\mathrm{Cov}(y_i,y_j) = Ey_iy_j - (Ey_i)(Ey_j) = Ee^{x_i+x_j} - (Ey_i)(Ey_j).$$

令 $a=e_i+e_j$，e_i 是第 i 个坐标为 1，其余坐标均为 0 的向量，则 $x_i+x_j=a'x$，

$$Ee^{x_i+x_j}=\int\cdots\int\left(\frac{1}{\sqrt{2\pi}}\right)^n\left|\sum\right|^{-\frac{1}{2}}e^{-\frac{1}{2}\left[(x-\mu)'\sum^{-1}(x-\mu)-2a'x\right]}dx$$

$$=\int\cdots\int\left(\frac{1}{\sqrt{2\pi}}\right)^n\left|\sum\right|^{-\frac{1}{2}}e^{-\frac{1}{2}Q}dx,$$

其中 $Q=(x-\mu-\sum a)'\sum^{-1}(x-\mu-\sum a)-a'\sum a-2\mu'a$，于是

$$\begin{aligned}Ey_iy_j&=Ee^{x_i+x_j}\\&=\exp\left\{a'\mu+\frac{1}{2}a'\sum a\right\}\\&=\exp\left\{\mu_i+\mu_j+\sigma_{ij}+\frac{1}{2}(\sigma_{ii}+\sigma_{jj})\right\},\end{aligned}$$

于是得

$$\mathrm{Cov}(y_i,y_j)=\left[\exp\left\{\mu_i+\mu_j+\frac{1}{2}(\sigma_{ii}+\sigma_{jj})\right\}\right][e^{\sigma_{ij}}-1].\tag{4.6}$$

实际上，从求 $\mathrm{Cov}(y_i,y_j)$ 的表达式时，我们已经证明了如下的公式

$$E\left(\prod_{i=1}^{n}y_i^{a_i}\right)=Ee^{a'x}=\exp\left\{a'\mu+\frac{1}{2}a'\sum a\right\},\tag{4.7}$$

从(4.7)式可以求出 y 的各阶混合乘积矩.

正的随机向量往往与实际问题中的基向量是吻合的，于是自然要问：如果基向量 ω 是遵从对数正态分布，则相应于 ω 的成分向量 x 会遵从什么样的分布？这里我们利用 §3 的结论来导出这一分布，下一章我们将用另一种方法来导出这一分布.

为了行文方便，考虑基向量

$$\underset{(n+1)\times 1}{\omega}=\begin{bmatrix}\omega_0\\\omega_1\\\vdots\\\omega_n\end{bmatrix},$$

令

$$t = \mathbb{1}'\omega = \sum_{i=0}^{n} \omega_i ,$$

相应的成分向量

$$\underset{(n+1)\times 1}{x} = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}, x_i = \frac{\omega_i}{t}, i = 0,1,2,\cdots,n.$$

由于 $x_i>0, \mathbb{1}'x=1$，因此成分向量 x 的分布实际上由其中 n 个分量完全决定. 为了方便，我们仍然称它为 x 的分布，联合分布密度是指后 n 个分量 $x_1,\cdots,x_n$ 的联合分布.

当 $\underset{(n+1)\times 1}{\omega} \sim \Lambda(\mu,\Sigma)$ 时，$\ln\omega \sim N(\mu,\Sigma)$，而

$$\ln x = \ln\omega - \mathbb{1}\ln t ,$$

令 $F=(-\mathbb{1} \quad I_n)$，$y=F\ln x=F\ln\omega-F\,\mathbb{1}\ln t=F\ln\omega$，于是 y 的分布密度由§3中的(3.6)式给出，因为 $\ln\omega$ 正态，$F\ln\omega$ 是 $\ln\omega$ 的线性函数，它还是正态分布，相应的密度是

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n |B_{11}|^{\frac{1}{2}} e^{-\frac{1}{2}(y-F\mu)'B_{11}(y-F\mu)}, \qquad (4.8)$$

其中

$$B_{11} = (I_n \quad 0)B\begin{pmatrix} I_n \\ 0 \end{pmatrix}, B = \Sigma^{-1} - \frac{\Sigma^{-1}\mathbb{1}\,\mathbb{1}'\Sigma^{-1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}.$$

由(4.8)式就可以导出 $x_1,\cdots,x_n$ 的联合密度，今后(4.8)式的密度是重要的，进一步的说明在下一章展开.

现在用(4.8)导 $x_1,\cdots,x_n$ 的联合密度. 由于 B 具有性质 $B\,\mathbb{1}=\Sigma^{-1}\mathbb{1}-\Sigma^{-1}\mathbb{1}\,\mathbb{1}'\Sigma^{-1}\mathbb{1}/\mathbb{1}'\Sigma^{-1}\mathbb{1}=0$，分块后就有

$$0 = B\,1 = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}\begin{pmatrix} \mathbb{1} \\ 1 \end{pmatrix} = \begin{pmatrix} B_{11}\,\mathbb{1} + B_{12} \\ B_{21}\,\mathbb{1} + B_{22} \end{pmatrix},$$

即

$$B_{21}' = B_{12} = -B_{11}\mathbb{1}, B_{22} = -B_{21}\,\mathbb{1} = \mathbb{1}'B_{11}\,\mathbb{1},$$

因此得

$$B = \begin{pmatrix} B_{11} & -B_{11}\mathbb{1} \\ -\mathbb{1}'B_{11} & \mathbb{1}'B_{11}\mathbb{1} \end{pmatrix}. \tag{4.9}$$

令

$$y_i = \ln x_i - \ln x_0, i = 1,2,\cdots,n,$$

于是

$$J(y_1,\cdots,y_n|x_1,\cdots,x_n) = \begin{vmatrix} x_1^{-1}+\frac{1}{x_0} & \frac{1}{x_0} & \cdots & \frac{1}{x_0} \\ \frac{1}{x_0} & x_2^{-1}+\frac{1}{x_0} & \cdots & \frac{1}{x_0} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{x_0} & \frac{1}{x_0} & \cdots & \frac{1}{x_n}+\frac{1}{x_0} \end{vmatrix}$$

$$= \prod_{i=0}^{n} x_i^{-1}.$$

而 $y_i = \ln x_i - \ln x_0$，写成向量时，即为 $y = F\ln x$，用引理 4.1 就得到 $x_1,\cdots,x_n$ 的联合密度是

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n |B_{11}|^{\frac{1}{2}}\left(\prod_{i=0}^{n} x_i^{-1}\right)e^{-\frac{1}{2}(F\ln x - F\mu)'B_{11}(F\ln x - F\mu)}. \tag{4.10}$$

将指数上二次型整理后，就有

$$\begin{aligned}&(F\ln x - F\mu)'B_{11}(F\ln x - F\ln\mu) \\ &= (\ln x - \mu)'F'B_{11}F(\ln x - \ln\mu),\end{aligned}$$

而

$$F'B_{11}F = \begin{pmatrix} I_n \\ -\mathbb{1}' \end{pmatrix} B_{11}(I_n - \mathbb{1})$$

$$= \begin{pmatrix} B_{11} & -B_{11}\mathbb{1} \\ -\mathbb{1}'B_{11} & \mathbb{1}'B_{11}\mathbb{1} \end{pmatrix}.$$

用(4.9)就知道 $F'B_{11}F = B$，这样可以将(4.10)改写为

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \mid B_{11} \mid^{\frac{1}{2}} \left(\prod_{i=0}^{n} x_i^{-1}\right) \mathrm{e}^{-\frac{1}{2}(\ln x-\mu)'B(\ln x-\mu)}, \quad (4.10)'$$

$$x > 0, \mathbb{1}'x = 1.$$

这就是 $x_1, \cdots, x_n$ 的联合密度,其中

$$B = \sum^{-1} - \frac{\sum^{-1} \mathbb{1}\mathbb{1}' \sum^{-1}}{\mathbb{1}' \sum^{-1} \mathbb{1}} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{matrix} n \\ 1 \end{matrix},$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\quad n \qquad 1$$

$|B_{11}|$ 的值可以证明它就是$(|\Sigma| \mathbb{1}'\Sigma^{-1}\mathbb{1})^{-1}$,这一结论留作练习.

§5. 狄氏分布

成分数据的统计分析中另一类重要的分布就是狄氏分布(Dirichlet distribution)及其推广的形式.狄氏分布与贝他函数、伽马函数有密切联系,它可以看成是多元的贝他函数,也可以很自然的从伽马分布导出.我们先扼要地介绍一下贝他函数、伽马函数、伽马分布,然后引出狄氏分布.

贝他函数 $B(a,b)$是由下式定义的:

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t, a > 0, b > 0, \quad (5.1)$$

伽马函数 $\Gamma(a)$是由下式定义的:

$$\Gamma(a) = \int_0^\infty t^{a-1}\mathrm{e}^{-t}\mathrm{d}t, a > 0, \quad (5.2)$$

它们之间有关系式

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \quad (5.3)$$

由 $B(a,b)$可以引出在(0,1)区间上的贝他分布,若 x 的分布密度为下述函数:

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1, \\ 0, & \text{其他}. \end{cases} \quad (5.4)$$

则称 x 遵从贝他分布，记为 $x\sim\beta(a,b)$，a 与 b 是分布中的两个参数.

由伽马函数可以引出在$(0,\infty)$上的伽马分布，若 x 的分布密度为下述函数：

$$f(x)=\begin{cases}\dfrac{b^a}{\Gamma(a)}x^{a-1}\mathrm{e}^{-bx}, & x>0,\\ 0, & \text{其他}.\end{cases} \tag{5.5}$$

则称 x 遵从伽马分布，记作 $x\sim\gamma(a,b)$，其中 a,b 均为正数，是分布的两个参数.

狄氏分布可以看成是贝他分布的推广，它可以由伽马分布引导而得.

令

$$B(a_1,a_2,\cdots,a_{n+1})$$
$$=\underset{\substack{x_i>0,i=1,2,\cdots,n\\ \sum_{i=1}^{n}x_i<1}}{\int\cdots\int}\left(\prod_{i=1}^{n}x_i^{a_i-1}\right)\left(1-\sum_{i=1}^{n}x_i\right)^{a_{n+1}-1}\mathrm{d}x_1\cdots\mathrm{d}x_n, \tag{5.6}$$

当 $n=1$ 时，$B(a_1,a_2)$就是(5.1)式的贝他函数，要注意(5.6)式右端是 n 重积分，$n=1$ 时就是(5.1)式右端的单积分.

可以用数学归纳法直接证明

$$B(a_1,a_2,\cdots,a_{n+1})=\prod_{i=1}^{n+1}\Gamma(a_i)\Big/\Gamma\left(\sum_{i=1}^{n+1}a_i\right), \tag{5.7}$$

这留作练习.下面我们用另一种方法证明这一等式，这一证明与成分数据能发生内在的联系.

假定基向量 $\omega'=(\omega_0,\omega_1,\cdots,\omega_n)$中各个分量是相互独立的；各自遵从伽马分布，参数有不相同的，因而不要求是同一分布. 即假定

$$\omega_i\sim\gamma(a_i,b),i=0,1,\cdots,n$$

且 $\omega_0,\omega_1,\cdots,\omega_n$ 相互独立. 于是令

$$t=\sum_{i=0}^{n}\omega_i,x_i=\omega_i/t,i=0,1,2,\cdots,n,$$

基向量 ω 相应的总量是 t，成分向量就是 x. 现在来求 t 与成分向量 x 的分布. 我们用引理 4.1 来导出 t 与 x 的联合密度，确切地说就是 $t, x_1, \cdots, x_n$ 的联合密度. 此时与引理 4.1 中相应的各个量是:

$D = R_{n+1}^{+}$,

$\widetilde{D} = \left\{(t, x_1, \cdots, x_n): t > 0, x_i > 0, i = 1, \cdots, n, \sum_{i=1}^{n} x_i < 1\right\}$,

$\omega_i = t x_i, i = 1, 2, \cdots, n$,

$\omega_0 = t\left(1 - \sum_{i=1}^{n} x_i\right) = t x_0$.

由于 ω_i 是相互独立的，它们的联合密度是

$$p(\omega_0, \cdots, \omega_n) = \left(\prod_{i=0}^{n} \frac{b^{a_i}}{\Gamma(a_i)} \omega_i^{a_i - 1}\right) e^{-b \sum_{i=0}^{n} \omega_i},$$

而 ω 对 $t, x_1, \cdots, x_n$ 的雅可比行列式为

$$J(\omega \mid t, x_1, \cdots, x_n) = \begin{vmatrix} 1 - \sum_{i=1}^{n} x_i & -t & \cdots & -t \\ x_1 & t & & 0 \\ \vdots & & \ddots & \\ x_n & 0 & & t \end{vmatrix} = t^n,$$

用引理 4.1 就得 $t, x_1, \cdots, x_n$ 的联合密度

$$\begin{aligned} f(t, x_1, \cdots, x_n) &= t^n p(t x_0, t x_1, \cdots, t x_n) \\ &= t^n \left[\prod_{i=0}^{n} \frac{b^{a_i}}{\Gamma(a_i)} (t x_i)^{a_i - 1}\right] e^{-b t \sum_{i=0}^{n} x_i}, \\ &\qquad (t, x_1, \cdots, x_n) \in \widetilde{D}, \end{aligned}$$

其中 $x_0 = 1 - \sum_{i=1}^{n} x_i$. 于是由 $\sum_{i=0}^{n} x_i = 1$，并令 $a = \sum_{i=0}^{n} a_i$ 后，就得

$$f(t, x_1, \cdots, x_n) = \begin{cases} b^a t^{a-1} e^{-bt} \prod_{i=0}^{n} \frac{x_i^{a_i - 1}}{\Gamma(a_i)}, & (t, x_1, \cdots, x_n) \in \widetilde{D}, \\ 0, & \text{其他}. \end{cases} \tag{5.8}$$

为了更简明突出密度函数的表达式，今后对密度函数只写出它的非零部分及相应的区域，于是(5.8)式可写成

$$f(t,x_1,\cdots,x_n)=b^a t^{a-1}e^{-bt}\prod_{i=0}^{n}\frac{x_i^{a_i-1}}{\Gamma(a_i)},(t,x_1,\cdots,x_n)\in\widetilde{D}.$$

将 $f(t,x_1,\cdots,x_n)$ 对 t 积分，利用公式

$$\int_0^\infty t^{a-1}e^{-bt}dt=\frac{\Gamma(a)}{b^a},$$

就得 $x_1,\cdots,x_n$ 的联合密度为

$$f_x(x_1,\cdots,x_n)=\frac{\Gamma(a)}{\prod_{i=0}^{n}\Gamma(a_i)}\left(\prod_{i=1}^{n}x_i^{a_i-1}\right)\left(1-\sum_{i=1}^{n}x_i\right)^{a_0-1},$$

$$x_i>0,i=1,2,\cdots,n,\sum_{i=1}^{n}x_i<1. \tag{5.9}$$

由于(5.9)是分布密度，密度在非0值区域上的积分是1，这样由(5.9)式就得到

$$\underset{\substack{x_i>0,i=1,2,\cdots,n\\ \sum_{i=1}^{n}x_i<1}}{\int\cdots\int}\left(\prod_{i=1}^{n}x_i^{a_i-1}\right)\left(1-\sum_{i=1}^{n}x_i\right)^{a_0-1}dx=\frac{\prod_{i=0}^{n}\Gamma(a_i)}{\Gamma\left(\sum_{i=0}^{n}a_i\right)}.$$

这实际上就是(5.7)，只要比较一下(5.6)的积分与上式左端的积分就可以看出.

不仅如此，我们从 $(t,x_1,\cdots,x_n)$ 的联合密度可以看出 t 与 $x_1,\cdots,x_n$ 是相互独立的，而且 t 的分布是 $\gamma(a,b)$，将以上结论用定理的形式给一个总结.

定理5.1 假定 $\omega_0,\omega_1,\cdots,\omega_n$ 相互独立，且

$$\omega_i\sim\gamma(a_i,b)\quad i=0,1,\cdots,n,$$

则基向量 $\omega'=(\omega_0,\omega_1,\cdots,\omega_n)$ 相应的总量与成分有如下的性质：

(1)总量 t 与成分 x 相互独立；

(2) $t\sim\gamma(a,b)$，其中 $a=\sum_{i=0}^{n}a_i$；

(3)成分 x 的分布密度是(5.9)式给出的狄氏分布 $D(a_0,a_1,\cdots,a_n)$.

狄氏分布(5.9)很明显是贝他分布的多元情形,它是成分数据的一个重要的分布类型.下面我们来计算狄氏分布的各阶矩.

假定 $x'=(x_0,x_1,\cdots,x_n)$ 是一成分向量,它的分布是狄氏分布 $D(a_0,a_1,\cdots,a_n)$,即密度为

$$\frac{\Gamma\left(\sum_{i=0}^{n} a_i\right)}{\prod_{i=0}^{n}\Gamma(a_i)}\prod_{i=0}^{n} x_i^{a_i-1}, \quad \begin{array}{l} x_i > 0, i = 0,1,\cdots,n, \\ \sum_{i=0}^{n} x_i = 1. \end{array} \tag{5.10}$$

注意(5.9)与(5.10)是同一个内容,只是形式不同,(5.10)的方便之处是它对于成分 x 的各个分量是对称的,形式上更为简明.令 $a'=(a_0,a_1,\cdots,a_n)$后,(5.10)还可以写成

$$\frac{\Gamma(\mathbb{1}'a)}{\prod_{i=0}^{n}\Gamma(a_i)}\prod_{i=0}^{n} x_i^{a_i-1}, x > 0, \mathbb{1}'x = 1. \tag{5.11}$$

令

$$E\left(\prod_{i=0}^{n} x_i^{\alpha_i}\right)=\int\cdots\int_{x>0,\mathbb{1}'x=1}\frac{\Gamma(\mathbb{1}'a)}{\prod_{i=0}^{n}\Gamma(a_i)}\left(\prod_{i=0}^{n} x_i^{\alpha_i+a_i-1}\right)\mathrm{d}x_1\cdots\mathrm{d}x_n,$$

利用(5.7)就可得右端的积分值,记 $\alpha'=(\alpha_0,\alpha_1,\cdots,\alpha_n)$后,就有

$$E\left(\prod_{i=0}^{n} x_i^{\alpha_i}\right)=\left[\prod_{i=0}^{n}\frac{\Gamma(\alpha_i+a_i)}{\Gamma(a_i)}\right]\frac{\Gamma(\mathbb{1}'a)}{\Gamma(\mathbb{1}'a+\mathbb{1}'\alpha)}$$

因此,取 $\alpha_i=1$,就得一阶矩;$\alpha_i=2$,就得二阶矩.也即有:

(1)$Ex_i=\dfrac{a_i}{\mathbb{1}'a}, i=0,1,\cdots,n$;

(2)$Ex_ix_j=\dfrac{a_ia_j}{(\mathbb{1}'a+1)\mathbb{1}'a}, i\neq j, i,j=0,1,\cdots,n$;

(3)$Ex_i^2=\dfrac{a_i(a_i+1)}{(\mathbb{1}'a+1)\mathbb{1}'a}, i=0,1,\cdots,n$;

(4)$\mathrm{Var}(x_i)=a_i(\mathbb{1}'a-a_i)/[(\mathbb{1}'a)^2(\mathbb{1}'a+1)], i=0,1,\cdots,n$;

(5)当 $i \neq j$ 时,

$$\mathrm{Cov}(x_i, x_j) = -a_i a_j / [(\mathbb{1}'a)^2(\mathbb{1}'a + 1)].$$

值得注意的是这种求各阶矩的方法是一个标准的方法,它利用密度中的参数和密度在全空间上积分为1,不需计算积分可以直接表示出来.我们分别用伽马分布及二项分布为例说明这一点.

若随机变量 x 的密度为

$$\frac{b^2}{\Gamma(a)} x^{a-1} \mathrm{e}^{-bx}, x > 0.$$

于是由密度积分为1,就得等式

$$\int_0^\infty x^{a-1} \mathrm{e}^{-bx} \mathrm{d}x = \frac{\Gamma(a)}{b^a}, a > 0, b > 0.$$

要求 x 的 k 阶原点矩时,有

$$Ex^k = \frac{b^a}{\Gamma(a)} \int_0^\infty x^k \cdot x^{a-1} e^{-bx} \mathrm{d}x.$$

右端积分只是将参数 a 用 $a+k$ 来代,从上述积分等式,马上就知道它的值是 $\frac{\Gamma(a+k)}{b^{a+k}}$,这样就求出 $Ex^k = \frac{a(a+1)\cdots(a+k-1)}{b^k}$,用一个式子来表示时,就有

$$Ex^k = \frac{b^a}{\Gamma(a)} \int_0^\infty x^k x^{a-1} \mathrm{e}^{-bx} \mathrm{d}x$$

$$= \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+k)}{b^{a+k}} = \frac{1}{b^k} \prod_{i=1}^k (a + i - 1).$$

若 x 遵从二项分布,其密度为

$$\binom{n}{k} \theta^k (1-\theta)^{n-k}, k = 0, 1, \cdots, n$$

于是有熟知的等式

$$\sum_{k=0}^n \binom{n}{k} \theta^k (1-\theta)^{n-k} = (\theta + 1 - \theta)^n = 1.$$

要求 x 的 m 阶原点矩,就有

$$Ex^m = \sum_{k=0}^n k^m \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

这时似乎没有等式可用，但注意到

$$\begin{aligned} Ex(x-1) &= \sum_{k=0}^{n} k(k-1)\binom{n}{k}\theta^k(1-\theta)^{n-k} \\ &= \sum_{k=2}^{n} k(k-1)\binom{n}{k}\theta^k(1-\theta)^{n-k} \\ &= n(n-1)\theta^2\sum_{k=0}^{n-2}\binom{n-2}{k}\theta^k(1-\theta)^{n-k-2} \\ &= n(n-1)\theta^2. \end{aligned}$$

实际上用了对参数 $n-2$ 的密度之和为 1 的等式. 可以求出 $Ex(x-1)(x-2)$等各阶乘矩，从阶乘矩再导出原点矩和中心矩.

以上这两个例子很有代表性，但矩的公式不用密度的性质，也早已知道，因此往往忽略了这一标准的利用密度的方法，求狄氏分布的矩，实际上是用了这一方法，今后会常用这一标准方法，到时就不细写了.

由于贝他函数与狄氏分布的联系，因此从贝他分布的性质可以导出一些相应的狄氏分布的性质，下面举例来给以说明.

设 $x\sim\beta(a,b)$，即 x 的密度为

$(B(a,b))^{-1}x^{a-1}(1-x)^{b-1}, 1>x>0, a>0, b>0,$

取 $y=\frac{x}{1-x}$，于是 y 在$(0,\infty)$上有密度，注意到

$$dy=\frac{1-x+x}{(1-x)^2}dx=\frac{1}{(1-x)^2}dx,$$

且

$$1+y=\frac{1}{1-x}.$$

因此 y 的密度就是

$$(B(a,b))^{-1}y^{a-1}(1+y)^{-(a+b)},\quad y>0.$$

这一分布用 $IB(a,b)$来表示，称为逆贝他分布(inverse beta). 注意，它不是贝他分布变量之逆的分布.

与这一个相类似，我们可以知道，若随机变量 $y\sim IB(a,b)$，则令 $x=y/(1+y)$后，就知道 $x\sim\beta(a,b)$.

于是可以将这个结果推广到狄氏分布. 设 $x_1,\cdots,x_n$ 遵从狄氏分布 $\mathrm{D}(a_0,a_1,\cdots,a_n)$,考虑

$$y_i=\frac{x_i}{1-\sum_{j=1}^{n}x_j},\quad i=1,2,\cdots,n,$$

于是同样可以求得雅可比为

$$J(x_1,\cdots,x_n\mid y_1,\cdots,y_n)=\Big(1+\sum_{j=1}^{n}y_j\Big)^{-(n+1)}.$$

代入分布密度,就得 $y_1,\cdots,y_n$ 的联合密度为下式中的 $p(y_1,\cdots,y_n)$:

$$p(y_1,\cdots,y_n)=\frac{\Gamma\Big(\sum_{i=0}^{n}a_i\Big)}{\prod_{i=0}^{n}\Gamma(a_i)}\Big(\prod_{i=1}^{n}y_i^{a_i-1}\Big)$$

$$\times\Big(1+\sum_{j=1}^{n}y_j\Big)^{-\sum_{i=0}^{n}a_i}.$$

因此

$$\frac{\Gamma\Big(\sum_{i=0}^{n}a_i\Big)}{\prod_{i=0}^{n}\Gamma(a_i)}\Big(1+\sum_{j=1}^{n}y_j\Big)^{-\sum_{i=0}^{n}a_i}\prod_{j=1}^{n}y_j^{a_j-1},$$

$$y_j>0,j=1,2,\cdots,n$$

称为逆狄氏分布,记为 $\mathrm{ID}(a_0,a_1,\cdots,a_n)$. 同样要注意逆指的是什么样的变换.

习　题　一

1. 直接验证:当 Σ_{11}^{-1} 存在时,只要 Σ^{-1} 存在,则下式就是 Σ 的逆阵 Σ^{-1},即有

$$\Sigma^{-1}=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}^{-1}=\begin{pmatrix}\Sigma_{11}^{-1} & 0\\ 0 & 0\end{pmatrix}$$

$$+\begin{bmatrix}\Sigma_{11}^{-1}\Sigma_{12}\\ -I\end{bmatrix}\Sigma_{22\cdot1}^{-1}\left(\Sigma_{21}\Sigma_{11}^{-1}\quad -I\right),$$

其中

$$\Sigma_{22\cdot1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

2. 证明习题1后，回答下述问题：

(a)证明中是否需要假定Σ是一正定阵？或要假定Σ是对称阵？

(b)如果上题中Σ^{-1}存在，Σ_{22}有逆，则Σ^{-1}分块写出的公式是什么？

3. 已知$\begin{pmatrix}x\\y\end{pmatrix}\begin{matrix}p\\q\end{matrix}$是联合正态分布$N(\mu,\Sigma)$，其中

$$\mu=\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}\begin{matrix}p\\q\end{matrix},\Sigma=\begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}\begin{matrix}p\\q\end{matrix}.$$
$$\qquad\qquad\qquad\qquad p\qquad q$$

若已知x的边缘分布为$N(a,A)$，且又知y对x的条件分布为$N(b,B)$，用a,b,A,B写出x对y的条件分布密度.

4. 如果已知$y\sim N(\mu,\Sigma)$，y是$m\times1$的随机变量，求$z=Fy$的分布密度，其中$F=(I_{m-1}\quad -\mathbb{1})$是$(m-1)\times m$的矩阵.

5. 已知二次型$Q=\sum_{i,j=1}^{n}a_{ij}(x_i-x_j)^2$，将$Q$写成$x'Bx$时，求$B$的表达式(即将$B$用$a_{ij}$表示出来). 并考虑以下问题：

(a)B的秩是多少？

(b)$B\mathbb{1}=0$是否成立？

6. 证明下述结论：若B是$n\times n$的非负定矩阵，并且B的秩$rkB=n-1$，则$B\mathbb{1}=0$的充分必要条件是：存在$\Sigma>0$使$B=\Sigma^{-1}-\dfrac{\Sigma^{-1}\mathbb{1}\mathbb{1}'\Sigma^{-1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}$.

7. 在上一题中，条件不变，a是任一给定的非零向量，则$Ba=0$的充要条件是什么？

8. 若$B=\Sigma^{-1}-\dfrac{\Sigma^{-1}\mathbb{1}\mathbb{1}'\Sigma^{-1}}{\mathbb{1}'\Sigma^{-1}\mathbb{1}}$，$B$是$n\times n$的矩阵，分块写出$B=\begin{pmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{pmatrix}\begin{matrix}n-1\\1\end{matrix}$，求证$|B_{11}|=\left(|\Sigma|\left(\mathbb{1}'\Sigma^{-1}\mathbb{1}\right)\right)^{-1}$.
$$n-1\quad 1$$

9. 用数学归纳法证明：$(a_i>0,i=1,2,\cdots,n)$

$$\underset{\substack{x_i>0, i=1,2,\cdots,n \\ \sum_{i=1}^{n} x_i<1}}{\int\cdots\int} \left(\prod_{i=1}^{n} x_i^{a_i-1}\right)\left(1-\sum_{i=1}^{n} x_i\right)^{a_0-1} \mathrm{d}x_1\cdots\mathrm{d}x_n = \frac{\prod_{i=0}^{n}\Gamma(a_i)}{\Gamma\left(\sum_{i=0}^{n} a_i\right)}$$

10. 利用习题 9 的结论，证明：$(a_i>0, i=1,2,\cdots,n)$

$$\underset{x_i>0, i=1,2,\cdots,n}{\int\cdots\int} \left(\prod_{i=1}^{n} x_i^{a_i-1}\right) f\left(\sum_{i=1}^{n} x_i\right) \mathrm{d}x_1\cdots\mathrm{d}x_n$$

$$= \frac{\prod_{i=1}^{n}\Gamma(a_i)}{\Gamma(a)} \int_0^{\infty} t^{a-1} f(t)\mathrm{d}t,$$

其中

$$a = \sum_{i=1}^{n} a_i.$$

11. 将习题 9—10 的结论推广到 x_i 是正定矩阵的情形，相应的 Γ 函数也作进一步的推广.

12. 将习题 10 中的 $f(\sum x_i)$ 改为 $f(\sum x_i^2)$，积分区域改为全空间，是否还能保持相应的结论?

附录　反正态分布及其推广

反正态分布及其推广越来越受到统计界的关注，一个重要的原因是它有广泛的应用背景. 从 1957 年文献[3]发表后，有很长一段时间没人注意这一分布，到了 70 年代，逐渐有人关注、发现这类分布与许多领域的现象有关，这些内容汇集在文献[4]的第一章中，文献[5]第一次将这类分布与成分数据相联系. 我们这个附录只介绍与成分数据有关的这些内容.

现在考虑 $\xi_1,\cdots,\xi_n$ 相互独立，但分布不相同，$\xi_i \sim \mathrm{IN}(\mu_i, \mu_i/\beta)$，参数 β 是相同的. 从 § 3 我们知道 ξ_i 的密度可以写成

$$\mu_i\sqrt{\frac{\beta}{2\pi}} t_i^{-\frac{3}{2}} \mathrm{e}^{-\frac{\beta}{2t_i}(t_i-\mu_i)^2},$$

因此，$\xi_1,\xi_2,\cdots,\xi_n$ 的联合密度是

$$\left(\sqrt{\frac{\beta}{2\pi}}\right)^n\left(\prod_{i=1}^n \mu_i t_i^{-\frac{3}{2}}\right)\mathrm{e}^{-\frac{\beta}{2}\sum_{i=1}^n\frac{(t_i-\mu_i)^2}{t_i}}, \tag{A.1}$$

$$t_i>0,i=1,2,\cdots,n.$$

又已知（见 §3 定理 3.2 的系 2）$s=\xi_1+\cdots+\xi_n$ 的分布是 $\mathrm{IN}\left(\sum_{i=1}^n\mu_i,\sum_{i=1}^n\mu_i/\beta\right)$，因此 s 的密度是：

$$\mu_+\sqrt{\frac{\beta}{2\pi}}s^{-\frac{3}{2}}\mathrm{e}^{-\frac{\beta}{2s}(s-\mu_+)^2}, \tag{A.2}$$

其中

$$\mu_+=\sum_{i=1}^n\mu_i.$$

考虑成分向量

$$x_i=\xi_i/s,i=1,2,\cdots,n,$$

于是$(x_1,x_2,\cdots,x_{n-1},s)$的联合分布密度从(A.1)得到：

$$\left(\sqrt{\frac{\beta}{2\pi}}\right)^n s^{n-1}\left(\prod_{i=1}^n\mu_i(sx_i)^{-\frac{3}{2}}\right)\mathrm{e}^{-\frac{\beta}{2}\sum_{i=1}^n\frac{(sx_i-\mu_i)^2}{sx_i}}, \tag{A.3}$$

$$s>0,x_i>0,i=1,2,\cdots,n,\sum_{i=1}^n x_i=1.$$

稍加整理，就得上式为

$$\left(\frac{\sqrt{\beta}}{2\pi}\right)^n\left(\prod_{i=1}^n\mu_i x_i^{-\frac{3}{2}}\right)s^{-\frac{n+2}{2}}\mathrm{e}^{-\frac{\beta}{2}\left(\sum_{i=1}^n\frac{\mu_i^2}{x_i}s^{-1}+s\right)+\beta\mu_+}, \tag{A.4}$$

$$s>0,x_i>0,i=1,2,\cdots,n,\sum_{i=1}^n x_i=1.$$

上式对 s 积分，就可以导出成分$(x_1,\cdots,x_{n-1})$的联合密度，注意到(A.4)式对 s 积分的值与贝塞尔函数有关，所以讨论起来比较复杂. 另一种考虑是看 $x_1,\cdots,x_{n-1}$对 s 的条件密度，将(A.4)式除以(A.2)式，就得条件密度，这一密度也可以导出另一类成分数据的分布，注意此时 s 是常数，相应的密度形式比较简单，即为

$$c\left(\prod_{i=1}^{n} x_i^{-\frac{3}{2}}\right)e^{-\sum_{i=1}^{n}\frac{a_i}{x_i}}, \tag{A.5}$$

$$x_i > 0, i = 1,2,\cdots,n, \sum_{i=1}^{n} x_i = 1,$$

其中 c 是与 n 及 a_i 有关的一个正则化常数.

类似的方法,也可以从逆反正态分布导出成分数据的分布类型.

反正态分布的推广,就是文献[4]的内容,推广的一般形式是密度为

$$p(x;\alpha,a,b) = \frac{\left(\frac{b}{a}\right)^{\alpha/2}}{2K_\alpha(\sqrt{ab})}x^{\alpha-1}e^{-\frac{1}{2}(ax^{-1}+bx)}, \quad x>0, \tag{A.6}$$

其中 $a>0, b>0, \alpha\in R$, $K_\alpha(\cdot)$是贝塞尔函数.从(A.6)就知道有

$$\int_0^\infty x^{\alpha-1}e^{-\frac{1}{2}(ax^{-1}+bx)}dx = \frac{2K_\alpha(\sqrt{ab})}{(b/a)^{\alpha/2}}.$$

这一公式可以求出广义反正态分布随机变量各阶矩的表达式.

参 考 文 献

[1] 张尧庭、方开泰(1982,1997),多元统计分析引论,科学出版社.

[2] Aitchison, J(1986), The Statistical Analysis of Compositional Data, Chapman & Hall.

[3] Tweedie, M. C. K.(1957), Statistical properties of the inverse Gaussian distribution I. Ann. Math. Statist., 28, 362—377.

[4] Jorgensen, B.(1982), Statistical Properties of the Generalized Inverse Gaussian Distribution. Lecture Notes in Statistics, Vol. 9, Springer-Verlag, New York.

[5] Barndorff-Nielsen, O. E. and Jorgensen, B. (1991), Some parametric models on the simplex. Journal of Mul. Analysis, 39, 106—116.

第二章 单形上的分布

§1. 成分与总量的独立性

基向量 ω 分解为总量 t 与成分 x，x 只是反映了 ω 的各个分量在总量 t 中所占的比例. 当 x 与 t 独立时 x 的分布与 t 的值无关，这样对 x 单独进行统计分析是有意义的；如果 x 与 t 不独立，则对 x 单独进行分析的意义就大为减弱. 例如，矿石样品中金属含量的百分比随着矿石的重量而变化，这样的百分比数据就难以单独分析；又如，人的血液中各类细胞所含的比例若随血液的多少而改变，则在验血时就难以从抽到的很少一部分血液来断定体内的情况. 这些实际例子都假定了成分 x 与总量 t 是相互独立的，因此在数学上讨论什么时候 ω 才具有这种性质，这是有意义的.

假定基向量 $\omega>0$，$\omega=(\omega_0,\omega_1,\cdots,\omega_n)'$，它的联合分布密度是 $p(\omega)=p(\omega_0,\omega_1,\cdots,\omega_n)$，于是

$$t=\sum_{i=0}^{n}\omega_i=\mathbb{1}'\omega,\quad x=\omega t^{-1},$$

$$x=(x_0,x_1,\cdots,x_n)'.$$

已知 ω 的密度后，可以很方便导出 t 与 $x_1,\cdots,x_n$ 的联合密度. 为了方便，以后我们说成分向量 x 的密度就是指它 n 个分量 $x_1,\cdots,x_n$ 的联合密度，注意 $x_0=1-\sum_{i=1}^{n}x_i$.

引理 1.1 假定 $\omega=(\omega_0,\omega_1,\cdots,\omega_n)$ 的联合密度是 $p(\omega_0,\omega_1,\cdots,\omega_n)$，则相应的 t 与 x 的联合密度是

$$t^n p\Big(t\Big(1-\sum_{i=1}^{n}x_i\Big),tx_1,\cdots,tx_n\Big). \tag{1.1}$$

证明 ω 的区域是 R_{n+1}^{+}，t 与 $x_1,\cdots,x_n$ 的变化区域是 $\widetilde{D}=$

$\{(t,x_1,\cdots,x_n):t>0,x_i>0,i=1,2,\cdots,n,\ \sum_{i=1}^n x_i<1\}$,而变换 $\omega_i=tx_i,i=1,2,\cdots,n,\omega_0=t\left(1-\sum_{i=1}^n x_i\right)$是1-1变换.上一章第五节中已求得

$$J(\omega\mid t,x_1,\cdots,x_n)=t^n,$$

因此(1.1)式就是 t,x 的联合密度.

现在,由引理1.1导出基向量 ω 的密度使相应的 t 与 x 独立的充要条件.

定理 1.1 假定基向量 ω 的联合密度是函数 $p(\omega_0,\omega_1,\cdots,\omega_n)$,则有:$t$ 与 x 独立的充要条件是存在函数 $f(\cdot)$ 使得

$$p(c\omega_0,c\omega_1,\cdots,c\omega_n)=\frac{f\left(c\sum_{i=0}^n\omega_i\right)}{c^nf\left(\sum_{i=0}^n\omega_i\right)}p(\omega_0,\cdots,\omega_n)\qquad(1.2)$$

对任一 $c>0$ 都成立.

证明 充分性.若(1.2)式成立,由引理1.1得到 t 与 x 的联合密度为

$$\begin{aligned}&t^np(tx_0,tx_1,\cdots,tx_n)\\&=t^n\frac{f(t)}{t^nf(1)}p(x_0,x_1,\cdots,x_n)\\&=\left(\frac{1}{f(1)}\right)f(t)p\left(1-\sum_{i=1}^n x_i,x_1,\cdots,x_n\right).\end{aligned}$$

很明显,t 与 x 是相互独立的.

必要性.若 x 与 t 独立,相应的密度分别用 $h(x_1,\cdots,x_n)$ 与 $f(t)$ 表示.由引理1.1知道应有关系式

$$t^np(tx_0,tx_1,\cdots,tx_n)=f(t)h(x_1,\cdots,x_n),$$

即为

$$p(tx_0,tx_1,\cdots,tx_n)=\frac{f(t)}{t^n}h(x_1,\cdots,x_n).$$

用 $\omega_i=tx_i$ 代入上式,得

$$p(\omega_0,\omega_1,\cdots,\omega_n)=\frac{f\left(\sum_{i=0}^{n}\omega_i\right)}{\left(\sum_{i=0}^{n}\omega_i\right)^n}h\left(\frac{\omega_1}{\mathbb{1}'\omega},\cdots,\frac{\omega_n}{\mathbb{1}'\omega}\right).$$

于是有：$\forall c>0$,

$$\begin{aligned}p(c\omega_0,c\omega_1,\cdots,c\omega_n)&=\frac{f\left(c\sum_{i=0}^{n}\omega_i\right)}{c^n\left(\sum_{i=0}^{n}\omega_i\right)^n}h\left(\frac{\omega_1}{\mathbb{1}'\omega},\cdots,\frac{\omega_n}{\mathbb{1}'\omega}\right)\\&=\frac{f\left(c\sum_{i=0}^{n}\omega_i\right)}{c^nf\left(\sum_{i=0}^{n}\omega_i\right)}\frac{f\left(\sum_{i=0}^{n}\omega_i\right)}{\left(\sum_{i=0}^{n}\omega_i\right)^n}h\left(\frac{\omega_1}{\mathbb{1}'\omega},\cdots,\frac{\omega_n}{\mathbb{1}'\omega}\right)\\&=\frac{f\left(c\sum_{i=0}^{n}\omega_i\right)}{c^nf\left(\sum_{i=0}^{n}\omega_i\right)}p(\omega_0,\omega_1,\cdots,\omega_n).\end{aligned}$$

这就证明了定理 1.1.

实际上，不论基向量 ω 相应的 t 与 x 是否独立，只要知道了 ω 的密度 $p(\omega_0,\omega_1,\cdots,\omega_n)$，由引理 1.1 立即可知成分 x 的分布密度是

$$\int_0^\infty t^n p(tx_0,tx_1,\cdots,tx_n)\mathrm{d}t. \tag{1.3}$$

例 1.1 假定 ω 的各个分量独立，各自遵从广义的伽马分布，即

$$\omega_i\sim\frac{b\lambda_i^{a_i}}{\Gamma(a_i)}\omega_i^{a_ib-1}\mathrm{e}^{-\lambda_i\omega_i^b},i=0,1,\cdots,n,$$

$$a_i>0,b>0,\lambda_i>0.$$

于是

$$p(\omega_0,\omega_1,\cdots,\omega_n)=b^{n+1}\left(\prod_{i=0}^{n}\frac{\lambda_i^{a_i}}{\Gamma(a_i)}\omega_i^{a_ib-1}\right)\mathrm{e}^{-\sum_{i=0}^{n}\lambda_i\omega_i^b}.$$

用(1.3)就得 ω 相应的成分向量 x 的分布密度为

$$\int_0^\infty t^n b^{n+1}\left(\prod_{i=0}^n \frac{\lambda_i^{a_i}}{\Gamma(a_i)}(tx_i)^{a_i b-1}\right)\mathrm{e}^{-t^b\sum\limits_{i=0}^n \lambda_i x_i^b}\mathrm{d}t$$

$$= b^{n+1}\left(\prod_{i=0}^n \frac{\lambda_i^{a_i}}{\Gamma(a_i)}x_i^{a_i b-1}\right)\int_0^\infty t^{ab-1}\mathrm{e}^{-t^b\left(\sum\limits_{i=0}^n \lambda_i x_i^b\right)}\mathrm{d}t,$$

其中 $a=\sum\limits_{i=0}^n a_i$. 利用公式

$$\int_0^\infty t^{ab-1}\mathrm{e}^{-\beta t^b}\mathrm{d}t=\frac{\Gamma(a)}{b\beta^a},$$

就得 $x_1,\cdots,x_n$ 的联合密度为

$$b^n\Gamma(a)\left(\prod_{i=0}^n \frac{\lambda_i^{a_i}}{\Gamma(a_i)}x_i^{a_i b-1}\right)\left(\sum_{i=0}^n \lambda_i x_i^b\right)^{-a}, \qquad (1.4)$$

注意

$$a=\sum_{i=0}^n a_i,\ x_0=1-\sum_{i=1}^n x_i,$$

$$x_i>0,\ i=1,2,\cdots,n,\ \sum_{i=1}^n x_i<1,$$

当 $b=1,\lambda_i=1,i=0,1,\cdots,n$ 时,它就是狄氏分布.

然而不论 ω 相应的 t 是否与 x 独立,x 的边缘分布总是有的,记为 $h(x_1,\cdots,x_n)$,于是对 x 配上一个独立的正随机变量 t,就可导出另一个基向量的分布,而这个基向量相应的 t 与 x 是相互独立的. 这就告诉我们,想寻找成分向量的分布类时,只需考虑 t 与 x 独立的情形就可以了. 从引理 1.1 和定理 1.1 就知道,t 与 x 独立时,ω 的密度 p 与 t,x 的各自的密度 f,h 有关系式

$$p(\omega_0,\omega_1,\cdots,\omega_n)=\frac{f(\mathbb{1}'\omega)}{(\mathbb{1}'\omega)^n}h\left(\frac{\omega_1}{\mathbb{1}'\omega},\cdots,\frac{\omega_n}{\mathbb{1}'\omega}\right),$$

即有

$$\ln p(\omega)=\ln g(\mathbb{1}'\omega)+\ln h\left(\frac{\omega_1}{\mathbb{1}'\omega},\cdots,\frac{\omega_n}{\mathbb{1}'\omega}\right).$$

很明显,上式右端第二项,把它看成 $\omega_0,\cdots,\omega_n$ 的函数时,它是

$\omega_0,\cdots,\omega_n$ 的齐 0 次的函数，而右端第一项只是$\mathbb{1}'\omega$ 的函数，x 的密度是由第二项决定的. 这就使我们从 $\omega_0,\omega_1,\cdots,\omega_n$ 的齐 0 次函数中去寻求相应的分布.

很明显，$\left\{\frac{\omega_i}{\omega_j}: i\neq j, i,j=0,1,\cdots,n\right\}$的函数是 $\omega_0,\omega_1,\cdots,\omega_n$ 的齐 0 次函数. 当然，最简单的函数是多项式，然而以$\frac{\omega_i}{\omega_j}$作为自变量的多项式实际上不可能形成正随机向量 ω 的分布密度，若与指数函数 $e^{-1'\omega}$一起可以形成一种混合分布，这一讨论可参看本章的习题. 另一种考虑是把 $\frac{\omega_i}{\omega_j}$ 的对数作为自变量，考虑 $\ln h\left(\frac{\omega_1}{\mathbb{1}'\omega},\cdots,\frac{\omega_n}{\mathbb{1}'\omega}\right)$是这些 $\ln\frac{\omega_i}{\omega_j}$ 的多项式，最简单的多项式是一次的或二次的，这两种情况正好对应于狄氏分布与加法逻辑正态分布，现在来说明这一点.

1. $\ln h\left(\frac{\omega_1}{\mathbb{1}'\omega},\cdots,\frac{\omega_n}{\mathbb{1}'\omega}\right)=\sum_{i,j=0}^{n}a_{ij}\ln\frac{\omega_i}{\omega_j}$.

注意到 $\ln(\omega_i/\omega_j)=\ln\omega_i-\ln\omega_j$，所以稍加整理就可将 $\sum_{i,j}a_{ij}\ln\frac{\omega_i}{\omega_j}$ 写成 $\sum_{j=0}^{n}(b_j-1)\ln\omega_j$，于是当 $\ln g(\mathbb{1}'\omega)=-\mathbb{1}'\omega$ 时，ω 的密度就是

$$c\Big(\prod_{j=0}^{n}\omega_j^{b_j-1}\Big)e^{-1'\omega},\quad \omega>0,$$

其中 c 是一个常数，它相应的成分 x 的分布密度，正好是狄氏分布. 如果 $\ln g(\mathbb{1}'\omega)=-(\mathbb{1}'\omega)^s$，于是 ω 的分布密度就是

$$c_*\Big(\prod_{j=0}^{n}\omega_j^{b_j-1}\Big)e^{-(1'\omega)^s},\quad \omega>0 \tag{1.5}$$

其中 c_* 也是一个常数，它相应的成分 x 的分布密度是狄氏分布，这一类分布的一般性质，我们在本章 §3 详细讨论.

2. $\ln h\left(\frac{\omega_1}{\mathbb{1}'\omega},\cdots,\frac{\omega_n}{\mathbb{1}'\omega}\right)=\sum_{j=0}^{n}(a_j-1)\ln\omega_j+\sum_{i,j=0}^{n}b_{ij}\left(\ln\frac{\omega_i}{\omega_j}\right)^2$.

稍加整理,上式右端可以写成向量 $\ln\omega$ 的一次型与二次型之和,即有如下的形式:

$$(a-\mathbb{1})'\ln\omega+(\ln\omega)'A\ln\omega.$$

由习题一中的5—8题,我们知道此时 A 具有性质 $A\mathbb{1}=0$. 如果 A 的秩是 n,且 $-A\geqslant 0$,则存在正定阵 $\sum$ 使 A 表成

$$-A=\left[\sum^{-1}-\sum^{-1}\mathbb{1}\mathbb{1}'\sum^{-1}/\mathbb{1}'\sum^{-1}\mathbb{1}\right]/2.$$

这样当 $a=0$ 时,只要 ω 相应的 t 与 x 独立,t 的分布密度是 $g(t)$,ω 相应的联合密度是

$$K(\mathbb{1}'\omega)g(\mathbb{1}'\omega)\left(\prod_{i=0}^{n}\omega_i\right)^{-1}\mathrm{e}^{-\frac{1}{2}(\ln\omega)'\Omega(\ln\omega)},\omega>0,$$

其中

$$\Omega=\sum^{-1}-\sum^{-1}\mathbb{1}\mathbb{1}'\sum^{-1}/\mathbb{1}'\sum^{-1}\mathbb{1},$$

K 是一个常数. 它相应的成分向量 x 的分布密度是

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{n}\left|\sum\right|^{-\frac{1}{2}}\left(\mathbb{1}'\sum^{-1}\mathbb{1}\right)^{-\frac{1}{2}}\left(\prod_{i=0}^{n}x_i\right)^{-1}\mathrm{e}^{-\frac{1}{2}(\ln x)'\Omega(\ln x)},\quad(1.6)$$

$$x=(x_0,x_1,\cdots,x_n)',x_i>0,i=0,1,\cdots,n,$$

$$x_0=1-\sum_{i=1}^{n}x_i,$$

其中

$$\Omega=\sum^{-1}-\sum^{-1}\mathbb{1}\mathbb{1}'\sum^{-1}/\mathbb{1}'\sum^{-1}\mathbb{1}.$$

这正是加法逻辑正态分布.

这就告诉我们,狄氏分布与加法逻辑正态分布在一定意义下是成分向量的分布类中最简单的两种分布.

更广泛一些的讨论是涉及基向量 ω 相应的大小 $G(\omega)$ 与形状 $z_G(\omega)$ 的独立性. 在这里我们给出一个有关大小的一个重要的结论.

引理 1.2[1)] 设 $x>0,y>0$,它们都是正的随机变量,且 x/y

1)参看文献[2]p.291.

既与 x 独立，又与 y 独立，则随机变量 x/y 一定是退化的.

证明 x/y 与 x 独立，x/y 与 y 独立，并且 x,y 均为正随机变量，因此有：

$$\ln x - \ln y \text{ 既与 } \ln x \text{ 独立，又与 } \ln y \text{ 独立，}$$

这就得到相应的特征函数

$$\begin{aligned} Ee^{it\ln y} &= Ee^{it(\ln y - \ln x + \ln x)} \\ &= Ee^{it(\ln y - \ln x)} Ee^{it\ln x}. \end{aligned}$$

同样地有

$$Ee^{it\ln x} = Ee^{it(\ln x - \ln y)} Ee^{it\ln y}.$$

因此

$$Ee^{it(\ln y - \ln x)} Ee^{it(\ln x - \ln y)} = 1 \quad \forall t \in R_1$$

这就知道 $\ln \frac{x}{y} = c$ 概率为 1 地成立，即 $\frac{x}{y}$ 是退化的.

有了引理 1.2，我们可以得到有关形状和大小独立性方面的一个重要的结论，这就是

定理 1.2[1)] 设 ω 是正的随机向量，$G_i(\omega)$ 是大小，相应的形状向量是 $z_i(\omega)$，$i=1,2$. 如果 $z_1(\omega)$ 与 $G_1(\omega)$ 独立，$z_1(\omega)$ 又与 $G_2(\omega)$ 独立，则 $G_2(\omega)/G_1(\omega)$ 一定是退化的.

证明 我们只要证明 $G_2(\omega)/G_1(\omega)$ 是 $z_1(\omega)$ 的函数，则由引理 1.2，从 $z_1(\omega)$ 与 $G_i(\omega)$，$i=1,2$ 各自的独立性，知道条件满足，就得所要的结论. 令

$$G_2(z_1(\omega)) = G_2(\omega/G_1(\omega)) = G_2(\omega)/G_1(\omega),$$

这就证明了 $G_2(\omega)/G_1(\omega)$ 是 $z_1(\omega)$ 的函数.

定理 1.2 告诉我们，与形状向量独立的大小一定是线性相关的，它们的比值是一个常数. 给定基向量 ω 之后，成分 x 与总量 t 是一对形状与大小的随机变量. 定理 1.2 告诉我们成分 x 若与总量 t 相互独立，则 x 不会与别的大小独立，除非它是 t 的倍数. 从证明中还可以看出形状向量彼此是可以互相表示的(也可以参见第一章的定理 2.1)，这告诉我们只要有一个形状向量 $z_1(\omega)$ 与相应的大小 $G_1(\omega)$ 独立，则所有的形状向量都与 $G_1(\omega)$ 独立.

近年来，对大小、形状等统计量有了进一步的扩展，这些已超越了本书的范围，有兴趣的读者可以参阅文献[2].

现在再来讨论成分向量 x 与 t 的关系，x 与 t 不独立时会是什么情况呢？能否给出条件分布与条件密度？我们以基向量 ω 是对数正态的情形为例，来作一点说明.

设 $\underset{(n+1)\times 1}{\omega}$ 服从对数正态分布，即 $\ln\omega \sim N(\mu, \Omega)$. 此时 ω 的联合密度为

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} |\Omega|^{-\frac{1}{2}}\left(\prod_{i=0}^{n} \omega_i^{-1}\right)\exp\left\{-\frac{1}{2}(\ln\omega - \mu)'\Omega^{-1}(\ln\omega - \mu)\right\}.$$

于是由引理 1.1 知道，成分向量 x 与总量 t 的联合密度为

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} |\Omega|^{-\frac{1}{2}} t^{-1}\left(\prod_{i=0}^{n} x_i^{-1}\right)\exp\left\{-\frac{1}{2}Q\right\},$$

其中

$$Q = (\ln x + \mathbb{1}\ln t - \mu)'\Omega^{-1}(\ln x + \mathbb{1}\ln t - \mu).$$

令 $l = \ln t$，于是 $\mathrm{d}l = \frac{1}{t}\mathrm{d}t$，因此 x 与 l 的联合密度为

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} |\Omega|^{-\frac{1}{2}}\left(\prod_{i=0}^{n} x_i\right)^{-1}\exp\left\{-\frac{1}{2}Q_1\right\}$$

其中

$$Q_1 = (\ln x + \mathbb{1}l - \mu)'\Omega^{-1}(\ln x + \mathbb{1}l - \mu)$$

将 Q_1 略加整理，写成 l 的平方和的形式，即

$$\begin{aligned} Q_1 &= (\mathbb{1}'\Omega^{-1}\mathbb{1})l^2 + 2l\,\mathbb{1}'\Omega^{-1}(\ln x - \mu) \\ &\quad + (\ln x - \mu)'\Omega^{-1}(\ln x - \mu) \\ &= (\mathbb{1}'\Omega^{-1}\mathbb{1})\left[l + \frac{\mathbb{1}'\Omega^{-1}(\ln x - \mu)}{\mathbb{1}'\Omega^{-1}\mathbb{1}}\right]^2 \\ &\quad + (\ln x - \mu)'\left[\Omega^{-1} - \frac{\Omega^{-1}\mathbb{1}\,\mathbb{1}'\Omega^{-1}}{\mathbb{1}'\Omega^{-1}\mathbb{1}}\right](\ln x - \mu), \end{aligned}$$

于是明显看出，对 l 积分后，就得 x 的联合分布密度为

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{n}(|\Omega|(\mathbb{1}'\Omega^{-1}\mathbb{1}))^{-\frac{1}{2}}\left(\prod_{i=0}^{n} x_i\right)^{-1} e^{-\frac{1}{2}Q_2},$$

其中

$$Q_2 = (\ln x - \mu)'\left[\Omega^{-1} - \frac{\Omega^{-1}\mathbb{1}\mathbb{1}'\Omega^{-1}}{\mathbb{1}'\Omega^{-1}\mathbb{1}}\right](\ln x - \mu).$$

于是就求得 l 对 x 的条件密度

$$p(l|x) = \left(\frac{\mathbb{1}'\Omega^{-1}\mathbb{1}}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\mathbb{1}'\Omega^{-1}\mathbb{1}\left(l+\frac{\mathbb{1}'\Omega^{-1}(\ln x-\mu)}{\mathbb{1}'\Omega^{-1}\mathbb{1}}\right)^2}, \quad (1.7)$$

也即 l 对 x 的条件密度是正态分布，相应的条件期望与条件方差是

$$\begin{cases} E\{l|x\} = -(\mathbb{1}'\Omega^{-1}\mathbb{1})^{-1}\mathbb{1}'\Omega^{-1}(\ln x - \mu), \\ \mathrm{Var}(l|x) = (\mathbb{1}'\Omega^{-1}\mathbb{1})^{-1}. \end{cases} \quad (1.8)$$

值得注意的是(1.8)中条件期望与 x 的值有关，而条件方差却与 x 无关. 当基向量 ω 的分量相互独立时，$\ln\omega$ 的各分量也相互独立，Ω 是一对角矩阵，Ω^{-1}是一对角阵，它相应的主对角元素是各自的方差的倒数，通常也称为精度，于是(1.8)式的条件期望就可以看成 $-(\ln x - \mu)$按 $\ln\omega$ 的精度进行加权平均，它的统计意义是很明显的.

(1.8)式还告诉我们，基向量 ω 服从对数正态分布时，总量 t 与成分向量 x 是不可能独立的，因为 $E\{l|x\}$随 x 而变. 尽管我们求得了 l 对 x 的条件分布是正态，但我们还未能求出 l 的边缘分布.

§2. 逻辑正态分布

逻辑正态分布类是 Aitchison 书中主要的内容，它有许多优良的性质，也可以从不同的方式引导出来. 这一节我们将论述各种等价的定义，然后讨论有关的性质.

定义 2.1 假定成分向量 $x = (x_0, x_1, \cdots, x_n)'$ 的函数 $y_i = \ln\frac{x_i}{x_0}, i = 1, 2, \cdots, n$ 服从 n 维正态分布 $N(\mu, \Sigma)$，则称成分 x 服

从加法逻辑正态分布.

这一名称的由来是因它与逻辑变换有关.对成分向量 x,将它用 $\ln\frac{x_i}{x_0}, i=1,2,\cdots,n$ 变换为 y_i 后,y_i 的变化范围是$(-\infty,\infty)$,将 x_i 用 y_i 表示时,就有

$$\begin{cases} x_i = e^{y_i}\Big/\left(1+\sum_{j=1}^{n} e^{y_j}\right), \qquad i=1,2,\cdots,n, \\ x_0 = \left(1+\sum_{j=1}^{n} e^{y_j}\right)^{-1}. \end{cases} \tag{2.1}$$

逻辑变换是英文 logit 的音译,y_i 是通过对 x_i 取对数(log − it)这样变换得来的.

定义 2.2 假定成分向量 $x=(x_0,x_1,\cdots,x_n)'$ 的函数 $y_i = \ln\frac{x_i}{1-\sum_{j=1}^{i} x_j}, i=1,2,\cdots,n$ 服从正态分布 $N(\mu,\Sigma)$,则称成分 x 服从乘法逻辑正态分布.

此时的 x_i 用 y_j 来表示时,有关系式

$$\begin{cases} x_i = e^{y_i}\Big/\prod_{j=1}^{i}(1+e^{y_j}), \qquad i=1,2,\cdots,n, \\ x_0 = \left[\prod_{j=1}^{n}(1+e^{y_j})\right]^{-1}. \end{cases} \tag{2.2}$$

比较(2.1)、(2.2)表达式中 x_i 的分母,一个是连加的形式,一个是连乘的形式,所以一个称为加法逻辑正态分布,一个称为乘法逻辑正态分布.

现在来引入分割逻辑正态分布.若对成分向量 x 进行分割,分为 k 段,则

$$\underset{(n+1)\times 1}{x} = \begin{pmatrix} x_{(1)} \\ \vdots \\ x_{(k)} \end{pmatrix} \begin{matrix} n_1 \\ \vdots \\ n_k \end{matrix}, \qquad n+1=\sum_{\alpha=1}^{k} n_\alpha .$$

令 $t_\alpha = \mathbb{1}'x_{(\alpha)}$,于是

$$t_1 + \cdots + t_k = \sum_{i=0}^{n} x_i = 1,$$

所以向量 $t = (t_1, \cdots, t_k)'$ 也是一个成分向量，记

$$s_{(\alpha)} = x_{(\alpha)}(\mathbb{1}'x_{(\alpha)})^{-1} = x_{(\alpha)}t_\alpha^{-1}, \alpha = 1,2,\cdots,k,$$

于是 $s_{(1)}, \cdots, s_{(k)}$ 就是 x 分割为 $x_{(1)}, \cdots, x_{(k)}$ 之后相应的子成分. 一个 k 分割相应地有

$$\underset{k\times 1}{t}, \underset{n_1\times 1}{s_{(1)}}, \cdots, \underset{n_k\times 1}{s_{(k)}},$$

这 $k+1$ 个向量，每个向量都是一个成分向量，各自的分布都可以由对数比来确定，于是有

定义 2.3 假定成分向量 x 对给定的 k 分割，相应于 t，$s_{(1)}, \cdots, s_{(k)}$，而 $t, s_{(1)}, \cdots, s_{(k)}$ 分别作加法或乘法逻辑变换后，相应的向量记为 $y_t, y_{(1)}, \cdots, y_{(k)}$. 如果这些 $y_t, y_{(1)}, \cdots, y_{(k)}$ 的联合分布是正态，则称成分 x 服从分割逻辑正态分布.

现在由这些定义来导出相应的 x 的分布密度.

1. 加法逻辑正态分布的密度

已知 $\underset{n\times 1}{y} \sim N(\mu, \Sigma)$，于是 y 的密度函数是

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right\}.$$

由于

$$y_i = \ln\frac{x_i}{x_0}, \quad i = 1,2,\cdots,n,$$

用向量写出就得

$$y = (-\mathbb{1} \quad I_n)\ln x, \quad \ln x = (\ln x_0, \ln x_1, \cdots, \ln x_n)'.$$

已知雅可比

$$J(y|x) = \prod_{j=0}^{n} x_j^{-1},$$

于是 x 的分布密度是

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n |\Sigma|^{-\frac{1}{2}} \left(\prod_{j=0}^{n} x_j\right)^{-1} e^{-\frac{1}{2}Q}, \tag{2.3}$$

其中

$$Q=(F\ln x-\mu)'\sum^{-1}(F\ln x-\mu), F=(-\mathbb{1}\quad I_n).$$

由于

$$F'=\begin{pmatrix}-\mathbb{1}'\\ I_n\end{pmatrix},\quad (FF')^{-1}=(I_n+\mathbb{1}\mathbb{1}')^{-1},$$

于是 $FF'(I_n+\mathbb{1}\mathbb{1}')^{-1}=I$，因而记 F^{+}[1] $=F'(FF')^{-1}$之后，

$$\begin{aligned}Q&=(F\ln x-FF^{+}\mu)'\sum^{-1}(F\ln x-FF^{+}\mu)\\&=(\ln x-F^{+}\mu)'F'\sum^{-1}F(\ln x-F^{+}\mu).\end{aligned}$$

注意到

$$\begin{aligned}F'\sum^{-1}F&=\begin{pmatrix}-\mathbb{1}'\\ I_n\end{pmatrix}\sum^{-1}(-\mathbb{1}\quad I_n)\\&=\begin{pmatrix}\mathbb{1}'\sum^{-1}\mathbb{1} & -\mathbb{1}'\sum^{-1}\\ -\sum^{-1}\mathbb{1} & \sum^{-1}\end{pmatrix},\end{aligned}$$

它具有性质

$$F'\sum^{-1}F\mathbb{1}=F'\sum^{-1}0=0,$$

再注意到 F'是满列秩的矩阵，于是 $F'\Sigma^{-1}F$ 的秩是 n，从上一章的习题就知道存在正定阵 Ω 使

$$F'\sum^{-1}F=\Omega^{-1}-\Omega^{-1}\mathbb{1}\mathbb{1}'\Omega^{-1}/\mathbb{1}'\Omega^{-1}\mathbb{1}.$$

这就与我们前面用其他方式导出的分布相同.

2. 乘法逻辑正态分布的密度

此时已知 $\underset{n\times 1}{y}\sim N(\mu,\Sigma)$，$y$ 的密度函数是

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n\left|\sum\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(y-\mu)'\sum^{-1}(y-\mu)\right\},$$

---

1) F^{+}有定义，我们这里既不引入定义，也不去证明 $F'(FF')^{-1}$就是 F^{+}.

注意到 $y_i = \ln x_i - \ln\left(1 - \sum_{j=1}^{i} x_j\right), i = 1,2,\cdots,n$,此时相应的雅可比行列式

$$J(y|x_1,\cdots,x_n) = \prod_{j=1}^{n}\left[\frac{1}{x_j} + \frac{1}{1 - \sum_{\alpha=1}^{j} x_\alpha}\right] = \prod_{j=1}^{n}\frac{1 - \sum_{\alpha=1}^{j-1} x_\alpha}{x_j\left(1 - \sum_{\alpha=1}^{j} x_\alpha\right)}$$

$$= \prod_{j=0}^{n} x_j^{-1},$$

于是乘法逻辑正态分布的密度为

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \left|\sum\right|^{-\frac{1}{2}}\left(\prod_{j=0}^{n} x_j\right)^{-1}\exp\left\{-\frac{1}{2}Q\right\},$$

$$x_j > 0, j = 0,1,\cdots,n, \quad x_0 = 1 - \sum_{j=1}^{n} x_j, \tag{2.4}$$

其中

$$Q = \sum_{i,j=1}^{n}\sigma^{ij}\left(\ln x_i - \ln\left(1 - \sum_{\alpha=1}^{i} x_\alpha\right) - \mu_i\right)$$

$$\times\left(\ln x_j - \ln\left(1 - \sum_{\alpha=1}^{j} x_\alpha\right) - \mu_j\right),$$

σ^{ij}是$\sum^{-1}$中第(i,j)位置的元素.很明显,这一分布密度不是取对数后能化为 $\ln x$ 的一次型或二次型的函数了.

3. 分割逻辑正态分布的密度

这里以二分割、三分割为例,写出相应的分布密度,一般的情形留给读者自己写出.先看简单的二分割的情形.

将成分向量 x 分为二段 $\underset{(n+1)\times 1}{x} = \begin{bmatrix} x_{(1)} \\ x_{(2)} \end{bmatrix}\begin{matrix} n_1 \\ n_2 \end{matrix}$,此时有

$$t_\alpha = \mathbb{1}'x_{(\alpha)}, \quad s_{(\alpha)} = x_{(\alpha)}(\mathbb{1}'x_{(\alpha)})^{-1} = x_{(\alpha)}t_\alpha^{-1}, \quad \alpha = 1,2,$$

由于 $t_1 + t_2 = 1$,因此考虑一个变量 t_1 就够了,记 t_1 为 t,$s_{(\alpha)}$相应的对数比分别用 $y_{(\alpha)}, \alpha = 1,2$ 表示,即

$$x_{(1)} = (x_0, x_1, \cdots, x_{n_1-1})', x_{(2)} = (x_{n_1}, x_{n_1+1}, \cdots, x_n)',$$

$$\begin{cases} y_t = \ln \frac{t}{1-t}, \\ y_{(1)} = \left(\ln \frac{x_1}{x_0}, \cdots, \ln \frac{x_{n_1-1}}{x_0}\right)' = \left(\ln \frac{s_{11}}{s_{10}}, \cdots, \ln \frac{s_{1n_1-1}}{s_{10}}\right)', \\ y_{(2)} = \left(\ln \frac{x_{n_1+1}}{x_{n_1}}, \cdots, \ln \frac{x_n}{x_{n_1}}\right)' = \left(\ln \frac{s_{21}}{s_{20}}, \cdots, \ln \frac{s_{2n_2-1}}{s_{20}}\right)', \end{cases} \tag{2.5}$$

$y_t, y_{(1)}, y_{(2)}$的联合密度是$N(\mu, \Sigma)$，即密度函数是

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \left(|\Sigma|^{-\frac{1}{2}}\right) e^{-\frac{1}{2}Q}, \tag{2.6}$$

其中

$$Q = \left[(y_t, y'_{(1)}, y'_{(2)}) - \mu'\right] \Sigma^{-1} \left[\begin{pmatrix} y_t \\ y_{(1)} \\ y_{(2)} \end{pmatrix} - \mu\right].$$

将$y_t, y_{(1)}, y_{(2)}$分别用(2.5)相应的t及x_i代入(2.6)，注意到相应的雅可比行列式

$$J(y_t, y_{(1)}, y_{(2)} \mid t, s_{(1)}, s_{(2)}) = \left(t \prod_{j=0}^{n_1-1} s_{1j} \prod_{j=0}^{n_2-1} s_{2j}\right)^{-1}.$$

于是从(2.6)就可以得$t, s_{(1)}, s_{(2)}$的联合密度.

将成分向量x分为三段，

$$\underset{(n+1)\times 1}{x} = \begin{pmatrix} x_{(1)} \\ x_{(2)} \\ x_{(3)} \end{pmatrix} \begin{matrix} n_1 \\ n_2, \\ n_3 \end{matrix}$$

此时有

$$t_\alpha = \mathbb{1}' x_{(\alpha)}, \quad s_{(\alpha)} = x_{(\alpha)} (\mathbb{1}' x_{(\alpha)})^{-1} = x_{(\alpha)} t_\alpha^{-1}, \alpha = 1,2,3,$$

用t表示向量$\begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix}$，而$\mathbb{1}'t = t_1 + t_2 + t_3 = \mathbb{1}'x = 1$，因此$t$也是

一个成分向量.分别用 $y_t, y_{(1)}, y_{(2)}, y_{(3)}$ 表示 $t, s_{(1)}, s_{(2)}, s_{(3)}$ 所形成的对数比的向量，和上面二分割的情形相似，可以从 $y_t, y_{(1)}, y_{(2)}, y_{(3)}$ 的联合密度为正态 $N(\mu, \Sigma)$，导出 $t, s_{(1)}, s_{(2)}, s_{(3)}$ 的联合密度，注意这时相应的雅可比行列式是

$$J(y_t, y_{(1)}, y_{(2)}, y_{(3)} \mid t, s_{(1)}, s_{(2)}, s_{(3)}) = \prod_{\alpha=1}^{3} \left(t_\alpha \prod_{j=0}^{n_\alpha - 1} s_{\alpha j} \right)^{-1},$$

其中 $s_{\alpha j}, j = 0,1,2,\cdots,n_\alpha - 1, \alpha = 1,2,3$ 是 $s_{(\alpha)}$ 相应的各个分量.

从上面的讨论就可以导出一般 k 分割的情形，这一讨论留作习题.

由于加法逻辑正态分布具有相当的代表性，从定义 2.1 看出，这一定义对成分 x 的各个分量不具有对称性，其中 x_0 占有特殊地位，在讨论各种性质时不方便，为此我们引出等价的定义，然后利用等价的定义可以导出各种统计性质.我们用定理的形式给出等价的定义.

定理 2.1 成分向量 x 遵从加法逻辑正态分布的充分必要条件是：x 的任一组对数对比都服从正态分布.

证明 充分性是明显的.因为 $\underset{n\times 1}{y} = (-\mathbb{1} \quad I_n) \underset{(n+1)\times 1}{\ln x}$ 是一组特殊的对数对比，所以 y 服从正态分布.现在来证必要性.由于全部对比的系数是形成一个 n 维子空间，而 $F = \begin{pmatrix} -\mathbb{1}' \\ I_n \end{pmatrix}$ 的列向量是这个子空间的一组基，因此任意一组对比的系数矩阵 A，它的各列都是对比系数向量，均可表示为 F 的线性组合，即存在 B 使 $A = FB$.因此 $A'\ln x = B'F'\ln x = B'y$，而 y 是正态分布，它的任一组线性函数都是正态分布，这就证明了必要性.

定理 2.1 告诉我们，加法逻辑正态分布可以用 x 的任一组对数对比都服从正态分布来刻划，也可以用它作为定义.这可以与正态分布的定义相类比：y 是 $p\times 1$ 的随机向量，y 的任一线性函数 $a'y$ 服从正态，则称 y 服从多元正态分布.下面我们利用定理 2.1

来证明加法逻辑正态分布的一些性质.

假定成分向量 $x:(n+1)\times 1$ 服从加法逻辑正态分布,则 x 具有如下的性质:

(1) x 的全部对数比 $\left\{\ln\frac{x_i}{x_j}: i\neq j,\quad i,j=0,1,\cdots,n\right\}$ 是联合正态分布.

因为每一个对数比 $\ln\frac{x_i}{x_j}=\ln x_i-\ln x_j$ 都是 x 的一个对数对比,所以全部对数比是一组特殊的对数对比,根据定理 2.1,它们服从正态分布.

(2) x 的任一 k 分割 $x=\begin{pmatrix}x_{(1)}\\ \vdots\\ x_{(k)}\end{pmatrix}$,相应的子成分 $s_{(1)},\cdots,s_{(k)}$ 是各自遵从加法逻辑正态分布,它们的对数对比都遵从正态分布.

由于 $s_{(a)}$ 的任意一组对数对比都是 x 的一组对数对比,根据定理 2.1 就知道应是正态分布,因此 $s_{(a)}$ 的分布是加法逻辑正态分布. 对 $s_{(1)},\cdots,s_{(k)}$ 都选定各自的对数对比 $A'_1\ln s_{(1)},\cdots,A'_k\ln s_{(k)}$,于是有

$$\begin{pmatrix}A'_1\ln s_{(1)}\\ \vdots\\ A'_k\ln s_{(k)}\end{pmatrix}=\begin{pmatrix}A'_1\ln x_{(1)}\\ \vdots\\ A'_k\ln x_{(k)}\end{pmatrix}=\begin{pmatrix}A'_1 & 0 & \cdots & 0\\ 0 & A'_2 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & A'_k\end{pmatrix}\ln x.$$

它们仍然是 x 的一组对数对比,由定理 2.1 知道它们的联合分布是正态分布.

当然,我们要注意多元正态分布的任一组线性函数仍然是正态分布,这包含了协方差矩阵可以是退化的情形,上面的结论中也是如此,这一点在应用时需要注意.

现在来讨论如何描述加法逻辑正态分布参数以及相关的统计性质.

设 $x'=(x_0,x_1,\cdots,x_n)$ 是成分向量,记

$$(\ln x)' = (\ln x_0, \ln x_1, \cdots, \ln x_n),$$

于是(2.1)相应的变换也就是

$$y = (-\mathbb{1} \quad I_n)\ln x, \qquad (2.7)$$

y 的分布是正态 $N(\mu, \Sigma)$. 描述向量 x 的均值自然是 $(Ex)' = (Ex_0, Ex_1, \cdots, Ex_n)$,由于$\mathbb{1}'x = 1$ 总成立,因此$\mathbb{1}'Ex = 1$ 自然也成立. y 的期望值由正态分布的性质知道 $Ey = \mu$,于是由(2.7)式得

$$\mu = Ey = (-\mathbb{1} \quad I_n)E\ln x . \qquad (2.8)$$

从 $E\ln x$ 去求出 Ex 是可以的,但它与从(2.1)式直接求出的 Ex 是否一样? 从(2.1)式得

$$\begin{cases} Ex_i = Ee^{y_i}\Big/\Big(1 + \sum\limits_{j=1}^{n} e^{y_j}\Big), \quad i = 1,2,\cdots,n, \\ Ex_0 = E\Big(1 + \sum\limits_{j=1}^{n} e^{y_j}\Big)^{-1}. \end{cases} \qquad (2.9)$$

然而,如果要讨论的 x 的性质可以用 y 的性质来描述的话,我们就只需处理正态总体,情况就简单得多. 这一方面,成分数据呈现出它的特性,在期望值向量和协方差阵上有明显的差别. 对成分向量 x,它的期望值 Ex 自然是统计中的重要参数向量,它反映了各个指标变量的平均百分比,并且一定满足其和为 1,即$\mathbb{1}'Ex = 1$. 而 x 的方差协方差矩阵 $\mathrm{Var}(x)$却是一定退化的,并且由于 $x_i > 0, \sum x_i = 1$,这样 x_i 与 x_j 之间往往呈现负相关. 然而在实际问题中,例如在一些地质勘探资料中,某些金属元素有伴生现象,如果铜的含量增加,伴随着金的含量也增加,它们之间呈现正相关性. 很可能在成分向量中可以分成若干组,各组之间是负相关,而组内却有正相关. 这一类统计性质的探讨是重要的. 加法逻辑正态分布的特点是将 x 的分析转化为$y = (-\mathbb{1} \quad I)\ln x$ 的分析,y 是正态分布,但是 y 的协方差矩阵在多大程度上反映了 x 协方差矩阵的性质,如何反映,这些都成了成分数据统计分析中的重要问题. 很明显,y 的协方差阵与 $\ln x$ 的协方差阵有密切的联系,因为 y 是 $\ln x$ 的线性函数. 但是 $\ln x$ 的协方差阵与 x 的协方差阵有什么关系,这就比较复杂了. 因此对于加法逻辑正态分布来说,如果

我们对成分向量 x 的特性的了解是由 $\ln x$ 来反映的，那么我们关注的是 $\ln x$ 的协方差阵与 y 的协方差阵的关系，这给统计分析带来很多方便．下面我们就给出反映这些关系的表达式．

从加法逻辑正态分布的性质可以知道，下列各个 $\ln x$ 的线性函数都是正态分布，它们的协方差矩阵与 $\ln x$ 的协方差矩阵有密切的联系．这些函数是

$$\underset{n\times 1}{y}=(-\mathbb{1}\quad I_n)\underset{(n+1)\times 1}{\ln x},$$

$$\underset{(n+1)\times 1}{\xi}=\left(I_{n+1}-\frac{1}{n+1}\mathbb{1}\,\mathbb{1}'\right)\underset{(n+1)\times 1}{\ln x},$$

$$t_{ij}=\ln x_i-\ln x_j,\quad i\neq j,i,j=0,1,\cdots,n.$$

设 $\ln x$ 的协方差矩阵为 V，$V=(v_{ij})i,j=0,1,\cdots,n$，则有

$$\begin{cases}\underset{n\times n}{\sum}=\mathrm{Var}(y)=(-\mathbb{1}\quad I)V\begin{pmatrix}-\mathbb{1}'\\ I\end{pmatrix},\\ \underset{(n+1)\times(n+1)}{\Gamma}=\mathrm{Var}(\xi)=\left(I_{n+1}-\frac{1}{n+1}\mathbb{1}\,\mathbb{1}'\right)V\left(I_{n+1}-\frac{1}{n+1}\mathbb{1}\,\mathbb{1}'\right),\\ \underset{(n+1)\times(n+1)}{T}=(\mathrm{Var}(t_{ij}))=(\mathrm{Var}(\ln x_i-\ln x_j))\\ \qquad=(v_{ii}+v_{jj}-2v_{ij}).\end{cases}\tag{2.10}$$

记 $D(V)$为 V 的主对角元素组成的对角矩阵，则在(2.10)中的 T 可以写成

$$T=D(V)\mathbb{1}\,\mathbb{1}'+\mathbb{1}\,\mathbb{1}'D(V)-2V.$$

(2.10)中的$\sum$，Γ，T 是可以相互表示的，例如

$$-2\sum=(-\mathbb{1}\quad I)T\begin{pmatrix}-\mathbb{1}'\\ I\end{pmatrix},$$

$$-2\Gamma=\left(I_{n+1}-\frac{1}{n+1}\mathbb{1}\,\mathbb{1}'\right)T\left(I_{n+1}-\frac{1}{n+1}\mathbb{1}\,\mathbb{1}'\right).$$

类似的关系式就可以逐一写出，这就留给读者自己当作练习．

以上这些在文献[1]中是很强调的，一些统计性质可以用 T，$\sum$或 Γ 来表示，看哪一个方便，而且它们彼此均可以转换．从数据

处理来看，用 V 表示是最方便的，因为从成分样本的数据很容易求出 V；另一方面，从加法逻辑正态分布的定义知道，许多统计分析都涉及 Σ，因此 Σ 与 V 的关系是重要的. 实际上(2.10)已给出

$$\Sigma = (-\mathbb{1} \quad I)V\begin{pmatrix} -\mathbb{1}' \\ I \end{pmatrix},$$

这是今后会常常用到的. (2.10)是将 Σ, Γ, T 用 V 表示，任何一个可以用其他三个来表示，这些关系式，读者自己去推导，或参见文献[1]中的内容，我们这里就不一一写出了.

对于加法逻辑正态分布，讨论成分向量的相关性的结构主要是讨论它的对数向量 $\ln x$ 的关系，它有以下几种不同的相关性.

定义 2.4 成分向量 x 称为对数比不相关，如果

$$\mathrm{Cov}\left(\ln\frac{x_i}{x_j}, \ln\frac{x_k}{x_l}\right) = 0, \tag{2.11}$$

对 i, j, k, l 四数不同时都成立.

成分向量 x 称为对数对比不相关，如果对任何两个正交的对比系数向量 a 与 b，$a'\ln x$ 与 $b'\ln x$ 总不相关，即 $a'\mathbb{1}=0, b'\mathbb{1}=0$ 且 $a'b=0$，则有

$$\mathrm{Cov}(a'\ln x, b'\ln x) = 0. \tag{2.12}$$

成分向量 x 的对数对比的系数向量 a 是单位向量时，即 $a'\mathbb{1}=0$且 $a'a=1$，对数对比的方差都相等，即

$$\mathrm{Var}(a'\ln x) \doteq c \tag{2.13}$$

对 $a'\mathbb{1}=0, a'a=1$ 成立. 则称 x 的对数对比各向同性.

上述三种不同的描述成分向量 x 的相关性的概念是程度不同的相关，它们之间有一定的联系，又有差别. 现在来讨论这种联系.

注意到对数比 $\ln(x_i/x_j)$ 是 $\ln x_i - \ln x_j$，它是一个对数对比，而当 i, j, k, l 都不相同时，$\ln(x_i/x_j)$ 与 $\ln(x_k/x_l)$ 的对比系数一定是正交的. 因此成分向量 x 是对数对比不相关时，一定是对数比不相关的. 然而，由对数比不相关却不能导出对数对比不相关，这很容易举出例子来说明.

各向同性与对数对比有什么关系呢？用成分向量 x 的对数的协差阵 V 来表示，就可以明了.

引理 2.1 设成分向量 x 是对数比不相关的，当 $n \geqslant 4$ 时则一定有

$$\sum = D(a) + b(\mathbb{1}\mathbb{1}' - I_n), \tag{2.14}$$

其中

$$D(a) = \begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_n \end{bmatrix}.$$

证明 记 $\sum = (\sigma_{ij})$ 后，就有在 $i \neq j$ 时，

$$\begin{aligned} \sigma_{ij} &= \mathrm{Cov}(\ln x_i - \ln x_0, \ln x_j - \ln x_0) \\ &= \mathrm{Cov}(\ln x_i - \ln x_0, \ln x_j - \ln x_k + \ln x_k - \ln x_0) \\ &= \mathrm{Cov}\left(\ln \frac{x_i}{x_0}, \ln \frac{x_j}{x_k}\right) + \mathrm{Cov}\left(\ln \frac{x_i}{x_0}, \ln \frac{x_k}{x_0}\right) \end{aligned}$$

对任一 k 成立，选 $k \neq 0, i, j$，上式右端第一项为 0，这就得 $\sigma_{ij} = \sigma_{ik}$，因此可得非对角元素全都相等，这就证明了(2.14).

引理 2.2 成分向量 x 各向同性的充分必要条件是：存在常数 c 使

$$\Gamma = c\left(I - \frac{1}{n+1}\mathbb{1}\mathbb{1}'\right). \tag{2.15}$$

证明 当 $a'\mathbb{1} = 0$ 时，$a = \left(I - \frac{1}{n+1}\mathbb{1}\mathbb{1}'\right)a \triangleq P_* a$，因此 $a'a = 1$ 时自然有 $a'P'_* P_* a = a'P_* a = 1$.

充分性：

$$\begin{aligned} \mathrm{Var}(a'\ln x) &= \mathrm{Var}(a'P'_* \ln x) = a'\Gamma a = a'P'_* V P_* a \\ &= ca'P_* a = c. \end{aligned}$$

必要性：由于 $\Gamma = P_* V P_*$，因此 $\Gamma\mathbb{1} = 0$，因此 $\mathbb{1}$ 是 Γ 的特征向量，相应的特征值是 0. 又

$$c = \max_{\substack{a'a=1 \\ a'\mathbb{1}=0}} \frac{a'\Gamma a}{a'a} = \min_{\substack{a'a=1 \\ a'\mathbb{1}=0}} \frac{a'\Gamma a}{a'a},$$

因此 Γ 的其余的非零特征根均相等且为c,因此有正交阵$(G\ \mathbb{1})$使

$\Gamma=(G\quad \mathbb{1})(cI)\begin{pmatrix}\mathbb{1}'\\ G'\end{pmatrix}$,即

$$\Gamma = cGG' + 0\mathbb{1}\,\mathbb{1}' \quad 且 \underset{(n+1)\times n}{G}\mathbb{1} = 0, G'G = I_n.$$

由

$$I_{n+1} = \left(G \quad \mathbb{1}\frac{1}{\sqrt{n+1}}\right)\begin{bmatrix} G' \\ \mathbb{1}'\dfrac{1}{\sqrt{n+1}}\end{bmatrix} = GG' + \frac{1}{n+1}\mathbb{1}\,\mathbb{1}',$$

就得

$$\Gamma = cGG' = c\left(I - \frac{1}{n+1}\mathbb{1}\,\mathbb{1}'\right).$$

定理 2.2 成分向量 x 各向同性与对数对比不相关是等价的.

证明 由引理 2.2,从 x 各向同性导出对数对比不相关是显然的,现在来证明它的逆.

对数对比不相关时一定有对数比不相关,因此由引理 2.1 知道,

$$\sum = D(a) + b(\mathbb{1}\,\mathbb{1}' - I), \quad D(a) = \begin{bmatrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_n\end{bmatrix},$$

或写成

$$\sigma_{ij} = \delta_{ij}\alpha_i + \alpha_0, i,j = 1,2,\cdots,n.$$

于是有 $\alpha_0 = b, \alpha_i = a_i - b$,今

$$\alpha_i + \alpha_0 = \mathrm{Var}\left(\ln\frac{x_i}{x_0}\right) = \mathrm{Cov}\left(\ln\frac{x_i}{x_0}, \ln\frac{x_i}{x_0}\right)$$

$$= \mathrm{Cov}\left(\ln\frac{x_i}{x_k} + \ln\frac{x_k}{x_0}, \ln\frac{x_i}{x_j} + \ln\frac{x_j}{x_0}\right).$$

当 i,j,k 不相等,i,j,k 都大于 0 时,由对数比不相关,得

$$\alpha_i + \alpha_0 = \mathrm{Cov}\left(\ln\frac{x_i}{x_k}, \ln\frac{x_i}{x_j}\right) + \mathrm{Cov}\left(\ln\frac{x_k}{x_0}, \ln\frac{x_j}{x_0}\right)$$

$$=\mathrm{Cov}\left(\ln\frac{x_i}{x_k},\ln\frac{x_i}{x_j}\right)+\alpha_0,$$

因此得 $i\neq k$,

$$\begin{aligned}\alpha_i-\alpha_k&=\mathrm{Cov}\left(\ln\frac{x_i}{x_k},\ln\frac{x_i}{x_j}\right)-\mathrm{Cov}\left(\ln\frac{x_i}{x_k},\ln\frac{x_j}{x_k}\right)\\&=\mathrm{Cov}(\ln x_i-\ln x_k,\ -2\ln x_j+\ln x_i+\ln x_k)\\&=0,\end{aligned}$$

于是 $\alpha_1=\cdots=\alpha_n=\alpha$,即$\sum=\alpha I+\alpha_0\mathbb{1}\mathbb{1}'$,而已知

$$\Gamma=\left(I-\frac{1}{n+1}\mathbb{1}\mathbb{1}'\right)\sum\left(I-\frac{1}{n+1}\mathbb{1}\mathbb{1}'\right),$$

因此

$$\Gamma=\alpha\left(I-\frac{1}{n+1}\mathbb{1}\mathbb{1}'\right),$$

也即 x 是各向同性的.

§3. 广义狄氏分布

在第一章,我们证明了下述结果:若基向量的分量相互独立,各自遵从伽马分布,则它所相应的成分向量就遵从狄氏分布.是否狄氏分布往往是与基向量各分量的独立性有关呢?我们将会看到并非如此,即使基向量的各分量之间并不独立,可以是正相关,也可以是负相关,然而,它们所相应的成分向量仍然是服从狄氏分布.进一步分析还可以看到,如果基向量的分布是一类很广泛的分布时,它可以导出比狄氏分布更广泛的一类成分分布,我们称它们为广义狄氏分布.这一节主要是导出这些分布密度.

在本章§1的例1.1中已经给出了由广义伽马分布导出的成分分布(1.4),它的一个特殊情形就是狄氏分布.现在考虑一类更广泛的分布,先引用一个积分公式,这个公式在讨论分布时起着重要的作用,我们用一条引理来叙述,以便今后不断地引用它.

引理3.1 设 $f(x)$是一元非负可测函数,它使下述积分的右

端有意义，于是就有下列公式：

$$\int\limits_{R_{n+1}^{+}} \left(\prod_{i=0}^{n} x_i^{a_i-1}\right) f\left(\sum_{i=0}^{n} x_i\right) \mathrm{d}x_0 \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

$$= \frac{\prod_{i=0}^{n} \Gamma(a_i)}{\Gamma\left(\sum_{i=0}^{n} a_i\right)} \int_0^{\infty} t^{a-1} f(t) \mathrm{d}t, \tag{3.1}$$

其中

$$a = \sum_{i=0}^{n} a_i.$$

证明 注意到(3.1)左端被积函数是非负的，因此积分有意义. 作变换：

$$t = \sum_{i=0}^{n} x_i,$$

$$y_i = x_i t^{-1}, \quad i = 1, 2, \cdots, n,$$

于是记 $x = (x_0, x_1, \cdots, x_n)'$, $y = (y_1, y_2, \cdots, y_n)'$ 后，

$$x_i = t y_i, \quad i = 1, 2, \cdots, n,$$

$$x_0 = t\left(1 - \sum_{i=1}^{n} y_i\right).$$

因此雅可比行列式(参看本章定理 1.1 的证明)

$$J(x \mid y, t) = t^n,$$

$$R_{n+1}^{+} \longrightarrow \left\{(y_1, \cdots, y_n, t):\ t > 0,\ y_i > 0,\quad \sum_{i=1}^{n} y_i < 1\right\},$$

于是(3.1)中左端积分

$$\int\limits_{R_{n+1}^{+}} \left(\prod_{i=0}^{n} x_i^{a_i-1}\right) f\left(\sum_{i=0}^{n} x_i\right) \mathrm{d}x_0 \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

$$= \int\limits_{y_i > 0, \sum_{i=1}^{n} y_i < 1} \left(\prod_{i=1}^{n} y_i^{a_i-1}\right) \left(1 - \sum_{i=1}^{n} y_i\right)^{a_0-1} \mathrm{d}y_1 \cdots \mathrm{d}y_n \int_0^{\infty} t^{a-1} f(t) \mathrm{d}t.$$

已知上式右端第一个积分为

$$\frac{\prod_{i=0}^{n}\Gamma(a_i)}{\Gamma\left(\sum_{i=0}^{n}a_i\right)},$$

于是引理得证.

不难看出,(3.1)中积分的值只依赖于单积分 $\int_0^\infty t^{a-1}f(t)\mathrm{d}t$,只要这个单积分值有限,记它为 $\Gamma(a;f)$,于是从(3.1)式就知道

$$\frac{\Gamma(a)/\Gamma(a;f)}{\prod_{i=0}^{n}\Gamma(a_i)}\left(\prod_{i=0}^{n}x_i^{a_i-1}\right)f\left(\sum_{i=0}^{n}x_i\right) \tag{3.2}$$

就是 R_{n+1}^{+}上的密度函数,这就是说,(3.2)可以作为一类基向量的联合密度.而且从引理 3.1 马上可以看出,如果基向量 $\omega=(\omega_0,\omega_1,\cdots,\omega_n)'$的联合密度是(3.2)的话,那么它所相应的成分向量与总量一定相互独立,并且成分向量遵从狄氏分布.很明显;$f(t)=\mathrm{e}^{-t}$时,$\Gamma(a;f)=\Gamma(a)$,引理给出的结论,就是由独立的伽马分布导出相应的狄氏分布.

从(3.2)的函数形式还可以看出,选择 $f(\cdot)$就相当于选择了总量 t 的密度,因为总量 t 的密度就是

$$\frac{1}{\Gamma(a;f)}t^{a-1}f(t),\qquad t>0. \tag{3.3}$$

现在我们可以从(3.2)的密度,导出基向量各分量之间并不独立,但它相应的成分向量依然是狄氏分布.

例 3.1 选 $f(t)=\mathrm{e}^{-t^b}$,$b>0$ 是已知常数.此时

$$\Gamma(a;f)=\int_0^\infty t^{a-1}\mathrm{e}^{-t^b}\mathrm{d}t=\frac{\Gamma\left(\frac{a}{b}\right)}{b}.$$

于是用 $\omega=(\omega_0,\omega_1,\cdots,\omega_n)'$表示基向量,它相应的(3.2)密度函数为

$$\frac{\Gamma(a)b}{\Gamma\left(\frac{a}{b}\right)\prod_{i=0}^{n}\Gamma(a_i)}\left(\prod_{i=0}^{n}\omega_i^{a_i-1}\right)e^{-\left(\sum_{i=0}^{n}\omega_i\right)^b},$$

$$\omega_i > 0,\quad i = 0,1,\cdots,n. \tag{3.4}$$

(3.4)式的 ω_i 之间是不独立的,利用(3.4)式对参数 $a_i>0, b>0$ 都是密度这一性质,或利用公式(3.1),都可以求出基向量 ω_i 之间的混合矩,得到

$$\begin{aligned}
E\omega_0 &= \frac{b\Gamma(a)}{\Gamma\left(\frac{a}{b}\right)\prod_{i=0}^{n}\Gamma(a_i)}\int_{R_{n+1}^{+}}\left(\prod_{i=0}^{n}\omega_i^{a_i-1}\right)\omega_0 e^{-\left(\sum_{i=0}^{n}\omega_i\right)^b}d\omega_0\cdots d\omega_n \\
&= \frac{b\Gamma(a)}{\Gamma\left(\frac{a}{b}\right)\prod_{i=0}^{n}\Gamma(a_i)}\frac{\Gamma\left(\frac{a+1}{b}\right)\left[\prod_{i=1}^{n}\Gamma(a_i)\right]\Gamma(a_0+1)}{b\Gamma(a+1)} \\
&= \frac{\Gamma\left(\frac{a+1}{b}\right)\Gamma(a)\Gamma(a_0+1)}{\Gamma\left(\frac{a}{b}\right)\Gamma(a+1)\Gamma(a_0)} \\
&= \frac{a_0}{a}\frac{\Gamma\left(\frac{a+1}{b}\right)}{\Gamma\left(\frac{a}{b}\right)}.
\end{aligned}$$

完全类似的计算(留作习题),可以证明下列等式:

$$\begin{cases}
E\omega_i = \dfrac{a_i}{a}\dfrac{\Gamma\left(\frac{a+1}{b}\right)}{\Gamma\left(\frac{a}{b}\right)}, & i = 0,1,\cdots,n, \\[2ex]
E\omega_i\omega_j = \dfrac{a_i a_j}{a(a+1)}\dfrac{\Gamma\left(\frac{a+2}{b}\right)}{\Gamma\left(\frac{a}{b}\right)}, & i \neq j, \\[2ex]
E\omega_i^2 = \dfrac{a_i(a_i+1)}{a(a+1)}\dfrac{\Gamma\left(\frac{a+2}{b}\right)}{\Gamma\left(\frac{a}{b}\right)}, & i = 0,1,\cdots,n.
\end{cases} \tag{3.5}$$

于是，可以求得 ω_i 与 $\omega_j(i\neq j)$ 的协方差

$$\begin{aligned}\mathrm{Cov}(\omega_i,\omega_j) &= E\omega_i\omega_j-(E\omega_i)(E\omega_j)\\ &=\frac{a_ia_j}{a(a+1)}\frac{\Gamma\left(\frac{a+2}{b}\right)}{\Gamma\left(\frac{a}{b}\right)}-\frac{a_ia_j}{a^2}\left[\frac{\Gamma\left(\frac{a+1}{b}\right)}{\Gamma\left(\frac{a}{b}\right)}\right]^2\\ &=\frac{a_ia_j}{a^2(a+1)\left[\Gamma\left(\frac{a}{b}\right)\right]^2}Q,\end{aligned}$$

其中

$$Q=a\Gamma\left(\frac{a}{b}\right)\Gamma\left(\frac{a+2}{b}\right)-(a+1)\left[\Gamma\left(\frac{a+1}{b}\right)\right]^2.$$

因此 ω_i 与 ω_j 不相关，正相关或负相关分别由 Q 的值是 0，是正还是负而定. 很明显，当 a 与 b 的值适当选取时，就可以出现正、负，或为 0 的 Q 值. 下面我们用数字来说明这一点：

当 $b=1$ 时，

$$\begin{aligned}Q&=a\Gamma(a)\Gamma(a+2)-(a+1)[\Gamma(a+1)]^2\\ &=a^2(a+1)[\Gamma(a)]^2-(a+1)a^2[\Gamma(a)]^2\\ &=0.\end{aligned}$$

当 $b=2$ 时，

$$\begin{aligned}Q&=a\Gamma(a/2)\Gamma\left(\frac{a}{2}+1\right)-(a+1)\left[\Gamma\left(\frac{a+1}{2}\right)\right]^2\\ &=\frac{a^2}{2}\left[\Gamma\left(\frac{a}{2}\right)\right]^2-(a+1)\left[\Gamma\left(\frac{a+1}{2}\right)\right]^2.\end{aligned}$$

对于 $a=1$ 时，$Q=\frac{1}{2}\left(\Gamma\left(\frac{1}{2}\right)\right)^2-2=\frac{\pi}{2}-2<0$，实际上只要 $0<a<2$，Q 均小于 0，此时 ω_i 与 ω_j 负相关.

当 $b=\frac{1}{2}$ 时，

$$\begin{aligned}Q&=a\Gamma(2a)\Gamma(2a+4)-(a+1)[\Gamma(2a+2)]^2\\ &=[\Gamma(2a)]^2[2a^2(2a+1)(2a+2)(2a+3)\\ &\quad-(a+1)4a^2(2a+1)^2]\end{aligned}$$

$$=[\Gamma(2a)]^2 4a^2(a+1)(2a+1)(2a+3-2a-1)$$
$$=8a^2(a+1)(2a+1)[\Gamma(2a)]^2>0,$$

对于 $a>0$，ω_i 与 ω_j 总是正相关.

从这个例子可以看出狄氏分布能反映基向量内部相关时，成分向量之间的联系，不限于基向量各分量彼此独立.

现在来考虑较一般的情形，导出广义的狄氏分布.

例 3.2 假定已知 $Q(\omega_0,\cdots,\omega_n)$ 是 $\omega_0,\cdots,\omega_n$ 的齐 s 次函数，即等式

$$Q(t\omega_0,\cdots,t\omega_n)=t^sQ(\omega_0,\cdots,\omega_n)$$

对 $t>0$ 均成立，并且

$$\int\limits_{R^+_{n+1}}\left(\prod_{i=0}^n\omega_i^{a_i-1}\right)\mathrm{e}^{-Q(\omega_0,\cdots,\omega_n)}\mathrm{d}\omega_0\mathrm{d}\omega_1\cdots\mathrm{d}\omega_n<\infty,$$

记上述积分值为 $c(\boldsymbol{a};Q)$，$\boldsymbol{a}=(a_0,a_1,\cdots,a_n)'$，于是

$$\frac{1}{c(\boldsymbol{a};Q)}\left(\prod_{i=0}^n\omega_i^{a_i-1}\right)\mathrm{e}^{-Q(\omega_0,\cdots,\omega_n)},\omega_i>0,i=0,1,\cdots,n.\tag{3.6}$$

就是 R^+_{n+1} 上的一个密度函数.

如果基向量 $\omega=(\omega_0,\cdots,\omega_n)'$ 遵从密度(3.6)，则成分向量 $x=(x_0,x_1,\cdots,x_n)'$ 与总量 t 的联合密度为

$$\frac{1}{c(\boldsymbol{a};Q)}\left(\prod_{i=1}^n x_i^{a_i-1}\right)\left(1-\sum_{i=1}^n x_i\right)^{a_0-1}t^{a-1}\mathrm{e}^{-t^sQ(x_0,\cdots,x_n)},$$
$$x_0=1-\sum_{i=1}^n x_i,\ a=\sum_{i=0}^n a_i,$$
$$t>0,\ x_i>0,\quad 1-\sum_{i=1}^n x_i>0.\tag{3.7}$$

于是将(3.7)对 t 积分后就得成分向量 x 的联合密度为

$$\frac{\Gamma\left(\frac{a}{s}\right)}{c(\boldsymbol{a};Q)}\frac{1}{s}\left(\prod_{i=1}^n x_i^{a_i-1}\right)\left(1-\sum_{i=1}^n x_i\right)^{a_0-1}[Q(x_0,x_1,\cdots,x_n)]^{-\frac{a}{s}}.\tag{3.8}$$

这一密度称为广义狄氏分布.

从(3.8)式可以看出,即使 $s=1$,此时密度为

$$\frac{\Gamma(a)}{c(\boldsymbol{a};Q)}\left(\prod_{i=1}^{n} x_i^{a_i-1}\right)\left(1-\sum_{i=1}^{n} x_i\right)^{a_0-1}\left[Q(x_0,x_1,\cdots,x_n)\right]^{-a}. \tag{3.9}$$

只有当

$$Q(x_0,x_1,\cdots,x_n)=\sum_{i=0}^{n} x_i=1$$

时,(3.9)才是狄氏分布.

很明显,选择不同的齐次函数 Q,就会导出不同的广义狄氏分布,如取

$$Q(\omega_0,\cdots,\omega_n)=\sum_{i=0}^{n}\lambda_i\omega_i^b,$$

(3.9)式就导出本章例 1.1 的分布.

尽管(3.8)式中有未知的 Q,但参数形成的常数项 $\Gamma\left(\frac{a}{s}\right)s^{-1}[c(\boldsymbol{a};Q)]^{-1}$ 的值依赖于向量 $\boldsymbol{a}$,s 与 Q 的形式,重要的是 $c(\boldsymbol{a};Q)$相应的积分能否表示出来,也就是说基向量 ω 的联合密度中的常数是否完全能用参数表示清楚.然而有意思的是,即使我们不知道 $c(\boldsymbol{a};Q)$的值,只要基向量 ω 遵从密度(3.6),ω_i 的各阶混合矩仍可以用$c(\boldsymbol{a};Q)$的值表示,这是标准的求矩方法又一次取得成功的例子.

设 $\omega=(\omega_0,\omega_1,\cdots,\omega_n)'$的联合密度是

$$\frac{1}{c(\boldsymbol{a};Q)}\left(\prod_{i=0}^{n}\omega_i^{a_i-1}\right)e^{-Q(\omega_0,\cdots,\omega_n)},\quad \omega_i>0,\quad i=0,1,\cdots,n.$$

于是

$$E\left(\prod_{i=0}^{n}\omega_i^{k_i}\right)=\frac{1}{c(\boldsymbol{a};Q)}\int_{R_{n+1}^{+}}\left(\prod_{i=0}^{n}\omega_i^{a_i+k_i-1}\right)e^{-Q}d\omega_0\cdots d\omega_n$$

$$=\frac{c(\boldsymbol{a}+\boldsymbol{k};Q)}{c(\boldsymbol{a};Q)},$$

其中

$$k = (k_0, k_1, \cdots, k_n)'.$$

如果 ω_i 是彼此独立的，$c(\boldsymbol{a}; Q)$就是一些密度中常数项乘积，它很方便就可以求出，所以从独立的 ω_i 导出相应的成分分布是不困难的．下面的几个例子将进一步说明这一点．

例 3.3 假定基向量 ω_i 相互独立，均遵从伽马分布，但参数均不相同，$(\omega_0, \cdots, \omega_n)$的联合密度为

$$\left(\prod_{i=0}^{n} \frac{\lambda_i^{a_i}}{\Gamma(a_i)} \omega_i^{a_i - 1}\right) e^{-\sum_{i=0}^{n} \lambda_i \omega_i}, \quad \omega_i > 0,\ i = 0, 1, 2, \cdots, n. \tag{3.10}$$

此时从例 3.2 就知道，它所相应的成分向量 $x = (x_0, x_1, \cdots, x_n)'$ 的分布是(参看(3.9)式)

$$\frac{\Gamma(a) \prod_{i=0}^{n} \lambda_i^{a_i - 1}}{\prod_{i=0}^{n} \Gamma(a_i)} \left(\prod_{i=0}^{n} x_i^{a_i - 1}\right) \left(\sum_{i=0}^{n} \lambda_i x_i\right)^{-a}, \tag{3.11}$$

其中

$$x_0 = 1 - \sum_{i=1}^{n} x_i, \quad a = \sum_{i=0}^{n} a_i.$$

当 λ_i 均相同时，(3.11)就成为狄氏分布．

容易看出，当基向量的 ω_i 相互独立，各自遵从逆伽马分布时，也可以导出类似的结论，这一内容留作习题．

例 3.4 假定基向量 $\omega = (\omega_0, \omega_1, \cdots, \omega_n)'$ 的联合分布密度为

$$\left(\prod_{i=0}^{n} \Gamma(a_i) \Gamma(b)\right)^{-1} \Gamma(a + b) \left(\prod_{i=0}^{n} \omega_i^{a_i - 1}\right) \left(1 + \sum_{i=0}^{n} \omega_i\right)^{-(a+b)}, \tag{3.12}$$

其中

$$a = \sum_{i=0}^{n} a_i, \quad a_i > 0, \quad i = 0, 1, \cdots, n,\ b > 0.$$

很明显，ω_i 之间是不独立的，然而它相应的成分向量的分布依然

是狄氏分布.

很容易看出,此时总量 t 与成分向量 x 是相互独立的,t 的分布是贝他分布.实际上,不独立的基向量的分量,它们的分布类型虽然不同,但往往仍可以导出相应的成分向量 x 的分布是狄氏分布.我们在习题中留下其他的例子,在正文中就不一一提及了.

从上面这些例子可以看出基向量的什么特性会导出成分向量的分布是狄氏分布.我们就以例 3.3 和 3.4 的情况来分析一下.

例 3.3 中,各个基分量 ω_i 是相互独立的,然而参数 λ_i 是否相同,决定了它们的成分向量是不是狄氏分布.例 3.4 中尽管基向量分量之间并不独立,但是它相应的成分向量仍然遵从狄氏分布.另一方面,从例 3.3 可以看出,基向量独立时,它相应的成分向量 x 是否服从狄氏分布取决于参数 λ_i,参数 λ_i 是一个尺度参数,当 ω_i 的度量单位发生变化时,λ_i 的值就会改变,适当调整尺度单位,就可以使 λ_i 的值彼此相同,相应的成分向量就是狄氏分布.从例 3.4 可以看出,尽管基向量的 ω_i 彼此是不独立的,a_i 也可以不相同,但只要密度函数中除了 $\prod_{i=0}^{n}\omega_i^{a_i-1}$ 这一项后是 $\sum_{i=0}^{n}\omega_i$ 的函数,那么总量 t 与成分向量一定相互独立,而且成分向量 x 的分布就是狄氏分布,狄氏分布的参数全由这些 a_i 来确定.这也反映了 ω_i 之间有一种均衡性,它们不均衡的项只反映在 $\prod_{i=0}^{n}\omega_i^{a_i-1}$ 这一项.从这些分析就可以看出,狄氏分布在某种程度上是由于基向量的某种均衡性决定的.

第一章狄氏分布的矩告诉我们,若成分向量$(x_0,x_1,\cdots,x_n)$的联合分布是狄氏分布,那么 x_i 的期望值 $Ex_i = a_i\Big/\left(\sum_{i=0}^{n}a_i\right)$, $i=0,1,\cdots,n$.这就告诉我们,成分 x_i 所占的"比重",只与 a_i 在 $\sum_{i=0}^{n}a_i$ 中的相对大小有关.进一步看,若基向量$(\omega_0,\omega_1,\cdots,\omega_n)$的联合密度是形如

$$\left(\prod_{i=0}^{n}\omega_i^{a_i-1}\right)f\left(\sum_{i=0}^{n}\omega_i\right)$$

这样的函数，那么 $f(t)$ 不论怎么变化，都只影响总量的分布，而不影响成分向量的分布，成分向量的分布只与上述函数中 $\prod_{i=0}^{n}\omega_i^{a_i-1}$ 这一项有关.

在结束本节前，我们再举一个例子，表明广义狄氏分布包含一些怎样的类型.

例 3.5 假定基向量 $\omega=(\omega_0,\omega_1,\cdots,\omega_n)'$ 的联合分布密度是

$$c\left(\prod_{i=0}^{n}\omega_i^{a_i-1}\right)\mathrm{e}^{-\left(\sum_{i=1}^{n}\omega_i+\frac{\omega_0^{n+1}}{\prod_{i=1}^{n}\omega_i}\right)},\quad \omega_i>0,\ i=0,1,\cdots,n, \tag{3.13}$$

其中 c 是一个常数，使上述函数在 R_{n+1}^{+} 上的积分为 1. 很明显，此时

$$Q(\omega_0,\omega_1,\cdots,\omega_n)=\sum_{i=1}^{n}\omega_i+\frac{\omega_0^{n+1}}{\prod_{i=1}^{n}\omega_i}$$

是 $\omega_0,\cdots,\omega_n$ 的齐一次函数，于是成分向量 $x_0,x_1,\cdots,x_n$ 的联合密度是

$$\begin{aligned}&c\Gamma(a)\left(\prod_{i=0}^{n}x_i^{a_i-1}\right)\left[\sum_{i=1}^{n}x_i+\frac{x_0^{n+1}}{\prod_{i=1}^{n}x_i}\right]^{-a},\\&x_i>0,\ i=1,2,\cdots,n,\\&x_0=1-\sum_{i=1}^{n}x_i>0,\ a=\sum_{i=0}^{n}a_i.\end{aligned} \tag{3.14}$$

要求出积分常数，就需要下面这个公式，用另一种方法求，可以参看文献[3]的第十四章的例题. 现在我们来证明有关(3.13)的一个积分公式，先作一些准备工作.

注意到(3.13)的积分常数 c 是一个积分

$$c^{-1} = \int\limits_{\substack{\omega_i>0 \\ i=0,1,2,\cdots,n}} \left(\prod_{i=0}^{n} \omega_i^{a_i-1}\right) e^{-\left(\sum_{i=1}^{n}\omega_i + \frac{\omega_0^{n+1}}{\prod_{i=1}^{n}\omega_i}\right)} d\omega_0 d\omega_1 \cdots d\omega_n. \tag{3.15}$$

作积分变量变换：

$$\omega_0 = \omega_0, \quad \omega_i = \omega_0 u_i, \quad u_i = \frac{\omega_i}{\omega_0}, \quad i = 1,2,\cdots,n,$$

于是

$$J(\omega_0,\omega_1,\cdots,\omega_n \mid \omega_0,u_1,\cdots,u_n) = \left(\prod_{i=1}^{n}\omega_0\right) = \omega_0^n,$$

因此(3.15)式可以写成：记 $a = \sum_{i=0}^{n} a_i$，

$$c^{-1} = \int\limits_{\substack{\omega_0>0 \\ u_i>0,\, i=1,2,\cdots,n}} \omega_0^{a-1}\left(\prod_{i=1}^{n} u_i^{a_i-1}\right) e^{-\omega_0\left(\sum_{i=1}^{n}u_i + \left(\prod_{i=1}^{n}u_i\right)^{-1}\right)} d\omega_0 du_1 \cdots du_n$$

$$= \Gamma(a) \int\limits_{u_i>0,\, i=1,2,\cdots,n} \left(\prod_{i=1}^{n} u_i^{a_i-1}\right)\left(\sum_{i=1}^{n} u_i + \left(\prod_{i=1}^{n} u_i\right)^{-1}\right)^{-a} du_1 \cdots du_n,$$

因此关键是求出上式右端的积分.

令

$$\int\limits_{u_i>0,\, i=1,2,\cdots,n} \left(\prod_{i=1}^{n} u_i^{a_i-1}\right)\left(\sum_{i=1}^{n} u_i + \left(\prod_{i=1}^{n} u_i\right)^{-1}\right)^{-a} du_1 \cdots du_n$$

$$= \int\limits_{u_i>0,\, i=1,2,\cdots,n} \left(\prod_{i=1}^{n} u_i^{a+a_i-1}\right)\left(\sum_{i=1}^{n}\left(\prod_{j=1}^{n} u_j\right) u_i + 1\right)^{-a} du_1 \cdots du_n. \tag{3.16}$$

引理 3.2

$$\int\limits_{u_i>0,\, i=1,3,\cdots,n} \left(\prod_{i=1}^{n} u_i^{b_i-1}\right) f\left(\sum_{i=1}^{n}\left(\prod_{j=1}^{n} u_j\right) u_i\right) du_1 du_2 \cdots du_n$$

$$=\frac{\prod_{i=1}^{n}\Gamma\left(b_i-\frac{b}{n+1}\right)}{(n+1)\Gamma(b/(n+1))}\int_0^{\infty}t^{\frac{b}{n+1}-1}f(t)\mathrm{d}t, \qquad (3.17)$$

其中

$$b=\sum_{i=1}^{n}b_i.$$

证明 对(3.17)左端积分作变量替换：令

$$z_i=u_i\Big(\prod_{j=1}^{n}u_j\Big),j=1,2,\cdots,n,$$

于是

$$\prod_{i=1}^{n}z_i=\Big(\prod_{j=1}^{n}u_j\Big)^{n+1},\quad \prod_{j=1}^{n}u_j=\Big(\prod_{i=1}^{n}z_i\Big)^{\frac{1}{n+1}},$$

$$u_i=z_i\Big(\prod_{j=1}^{n}z_j\Big)^{-\frac{1}{n+1}};$$

而且对 $i=1,2,\cdots,n$，有

$$\begin{aligned}\mathrm{d}z_i&=u_i\Big(\prod_{j\neq 1}u_j\Big)\mathrm{d}u_1+\Big(u_i\prod_{j\neq 2}u_j\Big)\mathrm{d}u_2+\cdots\\&\quad+\Big(2u_i\prod_{j\neq i}u_j\Big)\mathrm{d}u_i+\cdots+\Big(u_i\prod_{j\neq n}u_n\Big)\mathrm{d}u_n\\&=\Big(\prod_{j=1}^{n}u_j\Big)\sum_{j=1}^{n}\Big(\frac{u_i}{u_j}+\delta_{ij}\Big)\mathrm{d}u_j,\end{aligned}$$

因此

$$\begin{aligned}&J(z_1,\cdots,z_n\,|\,u_1,\cdots,u_n)\\&=\Big(\prod_{j=1}^{n}u_j\Big)^n\left|\frac{u_i}{u_j}+\delta_{ij}\right|\\&=\Big(\prod_{j=1}^{n}u_j\Big)^n\left|I+\begin{pmatrix}u_1\\ \vdots\\ u_n\end{pmatrix}(u_1^{-1}\cdots u_n^{-1})\right|\\&=(n+1)\Big(\prod_{j=1}^{n}u_j\Big)^n,\end{aligned}$$

这样

$$J(u_1,\cdots,u_n\,|\,z_1,\cdots,z_n)=(n+1)^{-1}\Big(\prod_{j=1}^{n}z_j\Big)^{-\frac{n}{n+1}}.$$

代入(3.17)左端,得左端积分为

$$(n+1)^{-1}\int\limits_{z_i>0,i=1,2,\cdots,n}\Big(\prod_{i=1}^{n}z_i^{b_i-1}\Big)\Big(\prod_{j=1}^{n}z_j\Big)^{-\frac{1}{n+1}(\sum b_j-n)-\frac{n}{n+1}}$$

$$\times f\Big(\sum_{j=1}^{n}z_j\Big)\mathrm{d}z_1\cdots\mathrm{d}z_n$$

$$=(n+1)^{-1}\int\limits_{z_i>0,i=1,2,\cdots,n}\Big(\prod_{i=1}^{n}z_i^{b_i-\frac{b}{n+1}-1}\Big)f\Big(\sum_{j=1}^{n}z_j\Big)\mathrm{d}z_1\cdots\mathrm{d}z_n,$$

再用引理 3.1 就得(3.17)右端.

将(3.17)用于(3.16),就得到

$$\frac{1}{\Gamma(a)}c^{-1}=\frac{\prod\limits_{i=1}^{n}\Gamma\Big(a_i+\dfrac{a_0}{n+1}\Big)}{(n+1)\Gamma\Big(a-\dfrac{a_0}{n+1}\Big)}\int_0^{\infty}t^{a-\frac{a_0}{n+1}-1}(1+t)^{-a}\mathrm{d}t$$

$$=\frac{\prod\limits_{i=1}^{n}\Gamma\Big(a_i+\dfrac{a_0}{n+1}\Big)}{(n+1)\Gamma\Big(a-\dfrac{a_0}{n+1}\Big)}\,\frac{\Gamma\Big(a-\dfrac{a_0}{n+1}\Big)\Gamma\Big(\dfrac{a_0}{n+1}\Big)}{\Gamma(a)}.$$

因此

$$c=\frac{(n+1)\Gamma(a)}{\Gamma(a)\Gamma\Big(\dfrac{a_0}{n+1}\Big)\prod\limits_{i=1}^{n}\Gamma\Big(a_i+\dfrac{a_0}{n+1}\Big)}.$$

代入(3.14),就得成分向量 $x_1,\cdots,x_n$ 的联合分布密度为

$$\frac{(n+1)\Gamma(a)}{\Gamma\Big(\dfrac{a_0}{n+1}\Big)\prod\limits_{i=1}^{n}\Gamma\Big(a_i+\dfrac{a_0}{n+1}\Big)}\Big(\prod_{i=0}^{n}x_i^{a_i-1}\Big)\left[\sum_{i=1}^{n}x_i+\frac{x_0^{n+1}}{\prod\limits_{i=1}^{n}x_i}\right]^{-a},$$

$$x_i>0,\quad x_0=1-\sum_{i=1}^{n}x_i,\quad i=0,1,2,\cdots,n,\quad a=\sum_{i=0}^{n}a_i.$$

§4. 其他成分分布

除了狄氏分布、广义狄氏分布、加法逻辑正态、乘法逻辑正态等这几类分布外，当然还可以列举一些其他的成分分布. 然而从本章前面讨论的情况可以看出，狄氏分布、逻辑正态分布无疑地是两类重要而基本的分布，它们是由相当广泛的一类基向量所导出的. 下面我们用举例的方式再给出一些成分分布，它们的统计性质与重要性看来和上述两类是不能相提并论的，我们列举一些只是表示成分分布还有一些别的类型.

原则上，在单形上具有非 0 值的密度函数都是一个成分分布，而且有很一般地导出成分分布的办法. 因为任何一个成分分布均可以看成是由基向量诱导而来的. 原则上诱导的方法可以有两个，现分别叙述如下，然后用一些例子来说明这两个方法是如何诱导的.

设基向量 $\omega=(\omega_0,\omega_1,\cdots,\omega_n)'$ 的联合密度是 $p(\omega_0,\omega_1,\cdots,\omega_n)$，从引理 1.1 知道，$\omega$ 相应的总量 $t=\mathbb{1}'\omega=\sum_{j=0}^{n}\omega_i$ 与 $x_i=\omega_i t^{-1}$, $i=0,1,\cdots,n$ 的联合密度是

$$t^n p(tx_0,tx_1,\cdots,tx_n),$$

$$t>0,\quad x_i>0,\quad i=0,1,\cdots,n,\quad \sum_{i=0}^{n}x_i=1.$$

将上述密度对 t 积分，就得成分向量的分布密度，这是一种方法，在前面我们已看到过. 另一个是考虑条件密度，$x_0,x_1,\cdots,x_n$ 对 t 的条件密度

$$p(x_0,x_1,\cdots,x_n\,|\,t)=t^n p(tx_0,tx_1,\cdots,tx_n)/g(t),$$

其中 $g(t)$是 t 的边缘密度. 如果让 $t=1$，代入上式，可见

$$p(x_0,x_1,\cdots,x_n\,|\,t=1)\propto p(x_0,x_1,\cdots,x_n).$$

这就告诉我们，任何一个基向量的分布密度 $p(\omega_0,\omega_1,\cdots,\omega_n)$，只要将 ω_i 换成成分 x_i，且不要作任何更动，就引导出一个成分分布

的核 $p(x_0,x_1,\cdots,x_n)$，只要在单形上对这个核积分求出正则化常数，就得完全的密度表达式. 也即有$(x_0,x_1,\cdots,x_n)$这个成分向量的密度是

$$f(x_0,x_1,\cdots,x_n)=\frac{p(x_0,x_1,\cdots,x_n)}{\displaystyle\int\limits_{\substack{x_i>0,i=0,1,\cdots,n\\ \sum_{i=0}^{n}x_i=1}} p(x_0,x_1,\cdots,x_n)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n}. \tag{4.1}$$

很明显，当 t 与成分向量 x 相互独立时，这两种方法给出的是相同的成分分布，狄氏分布由伽马分布导出的过程就是一个例子. 如果 t 与成分向量 x 不独立，给出的分布就不会相同.

例 4.1 反正态分布诱导的成分分布.

设基向量 $\omega_0,\omega_1,\cdots,\omega_n$ 的分量相互独立，各自遵从反正态分布：

$$\omega_i\sim\left(\frac{\lambda}{2\pi}\right)^{\frac{1}{2}}\omega_i^{-\frac{3}{2}}\mathrm{e}^{-\frac{\lambda}{2\mu_i^2}\left(\frac{\omega_i-\mu_i}{\sqrt{\omega_i}}\right)^2},\quad \omega_i>0,\quad i=0,1,2,\cdots,n.$$

于是$(\omega_0,\omega_1,\cdots,\omega_n)$的联合密度是

$$\begin{aligned}p(\omega_0,\cdots,\omega_n)&=\left(\frac{\lambda}{2\pi}\right)^{\frac{n+1}{2}}\left(\prod_{i=0}^{n}\omega_i^{-\frac{3}{2}}\right)\mathrm{e}^{-\frac{\lambda}{2}\sum_{i=0}^{n}\frac{(\omega_i-\mu_i)^2}{\mu_i^2\omega_i}}\\&=\left(\frac{\lambda}{2\pi}\right)^{\frac{n+1}{2}}\left(\prod_{i=0}^{n}\omega_i^{-\frac{3}{2}}\right)\mathrm{e}^{-\frac{\lambda}{2}\sum_{i=0}^{n}\left(\frac{\omega_i}{\mu_i^2}+\frac{1}{\omega_i}\right)+\lambda\sum_{i=0}^{n}\frac{1}{\mu_i}}.\end{aligned} \tag{4.2}$$

现在分别用两种方法来导出成分分布密度.

用(t,x)的联合分布，得到 x 的分布密度

$$\begin{aligned}f(x_1,\cdots,x_n)=&\left(\frac{\lambda}{2\pi}\right)^{\frac{n+1}{2}}\mathrm{e}^{\lambda\sum_{i=0}^{n}\frac{1}{\mu_i}}\left(\prod_{i=0}^{n}x_i\right)^{-\frac{3}{2}}\\&\times\int_0^{\infty}t^{-\left(\frac{n+1}{2}+1\right)}\mathrm{e}^{-\frac{\lambda}{2}\sum_{i=0}^{n}\left(\frac{tx_i}{\mu_i^2}+\frac{1}{tx_i}\right)}\mathrm{d}t\end{aligned} \tag{4.3}$$

上式右端中的积分与贝塞尔函数有关，利用贝塞尔函数

$$K_\beta(\sqrt{ab}) = \frac{1}{2}\left(\frac{a}{b}\right)^{\beta/2}\int_0^\infty x^{\beta-1}e^{-\frac{1}{2}(ax+bx^{-1})}dx, \qquad (4.4)$$

就可以将(4.3)式表示成 $K_\beta(\sqrt{ab})$ 的形式. 对(4.4)式右端作积分变换 $t = x^{-1}, dx = -t^{-2}dt$，于是有

$$K_\beta(\sqrt{ab}) = \frac{1}{2}\left(\frac{a}{b}\right)^{\beta/2}\int_0^\infty t^{-(\beta+1)}e^{-\frac{1}{2}(at^{-1}+bt)}dt, \qquad (4.5)$$

将它用于(4.3)的积分，

$$\beta = \frac{n+1}{2}, \quad a = \frac{\lambda}{2}\sum_{i=0}^{n}\frac{1}{x_i}, \quad b = \frac{\lambda}{2}\sum_{i=0}^{n}\frac{x_i}{\mu_i^2},$$

因此

$$f(x_1,\cdots,x_n) = \left(\frac{\lambda}{2\pi}\right)^{\frac{n+1}{2}}e^{\lambda\sum_{i=0}^{n}\frac{1}{\mu_i}}\left(\prod_{i=0}^{n}x_i\right)^{-\frac{3}{2}} \times \frac{2K_{\frac{n+1}{2}}\left\{\frac{\lambda}{2}\sqrt{\sum\frac{1}{x_i}\sum\frac{x_i}{\mu_i^2}}\right\}}{\left(\sum\frac{1}{x_i}\Big/\sum\frac{x_i}{\mu_i^2}\right)^{\frac{n+1}{2}}}, \qquad (4.6)$$

其中

$$x_0 = 1 - \sum_{i=1}^{n}x_i, \quad x_i > 0, \quad i = 0,1,2,\cdots,n.$$

这是用联合分布求积分的方法导出一种成分分布密度；用另一种条件分布密度的方法可以得到下面的另一种密度函数，它的核是

$$\left(\prod_{i=0}^{n}x_i\right)^{-\frac{3}{2}}e^{-\frac{\lambda}{2}\left(\sum_{i=0}^{n}\frac{x_i}{\mu_i^2}+\sum_{i=0}^{n}\frac{1}{x_i}\right)},$$

$$x_i > 0, \quad i = 0,1,\cdots,n, \quad x_0 = 1 - \sum_{i=1}^{n}x_i. \qquad (4.7)$$

这个函数在 $D_n = \left\{(x_1,\cdots,x_n): x_i > 0, \quad i = 1,2,\cdots,n, \sum_{i=1}^{n}x_i < 1\right\}$ 上积分，求出积分常数后，它就是一类成分分布密度，很明

显，这个积分是不易求出的．因为有 $x_0=1-\sum_{i=1}^{n}x_i$ 在指数的幂中，而且还有 $\left(1-\sum_{i=1}^{n}x_i\right)^{-1}$，这一类成分分布的密度和(4.6)明显地不同．类似地可以从广义反正态分布中诱导出别的分布．

例 4.2 从贝他分布导出成分分布．

设基向量$(\omega_0,\omega_1,\cdots,\omega_n)$的分量相互独立，各自遵从贝他分布，即

$$\omega_i\sim\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}\frac{\omega_i^{a_i-1}}{(1+\omega_i)^{a_i+b_i}},\omega_i>0,\quad i=0,1,2,\cdots,n.$$

用引理 1.1，就得(t,x)的联合密度为

$$t^n\prod_{i=0}^{n}\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}\frac{(tx_i)^{a_i-1}}{(1+tx_i)^{a_i+b_i}},\quad t>0,$$

$$x_0=1-\sum_{i=1}^{n}x_i,\quad x_i>0,\quad i=0,1,2,\cdots,n.$$

对 t 从 0 到∞积分，就得成分向量的密度：

$$\left(\prod_{i=0}^{n}\frac{\Gamma(a_i+b_i)}{\Gamma(a_i)\Gamma(b_i)}x_i^{a_i-1}\right)\int_0^{\infty}\frac{t^{a-1}\mathrm{d}t}{\prod_{i=0}^{n}(1+tx_i)^{a_i+b_i}}$$

其中

$$a=\sum_{i=0}^{n}a_i,\quad x_0=1-\sum_{i=1}^{n}x_i,\quad x_i>0,\quad i=0,1,\cdots,n.$$

例 4.3 设基向量$(\omega_0,\omega_1,\cdots,\omega_n)$的分布密度是

$$K\left(\prod_{i=0}^{n}\omega_i^{a_i-1}\right)\left(\sum_{i=0}^{n}b_i\omega_i\right)^{-q}\mathrm{e}^{-\sum_{i=0}^{n}\lambda_i\omega_i},$$

$$\omega_i>0,\ a_i>0,\ b_i>0,\ q>0,\ \lambda_i>0,\ i=0,1,2,\cdots,n,\tag{4.8}$$

其中 K 是一个常数，使这个函数积分为 1，因此先讨论 K 的值在什么条件下是一有限值．

利用伽马函数的性质：

$$\frac{\Gamma(q)}{\left(\sum_{i=0}^{n} b_i\omega_i\right)^q} = \int_0^\infty s^{q-1} e^{-s\left(\sum_{i=0}^{n} b_i\omega_i\right)} ds ,$$

于是

$$K^{-1} = \int_{\substack{\omega_i>0, i=0,1,\cdots,n \\ s>0}} \left(\prod_{i=0}^{n} \omega_i^{a_i-1}\right) e^{-\sum_{i=0}^{n}\lambda_i\omega_i - s\left(\sum_{i=0}^{n} b_i\omega_i\right)} \frac{s^{q-1}}{\Gamma(q)} d\omega_0\cdots d\omega_n ds$$

$$= \int_{\substack{\omega_i>0, i=0,1,\cdots,n \\ s>0,}} \frac{1}{\Gamma(q)} \left(\prod_{i=0}^{n} \omega_i^{a_i-1}\right) s^{q-1} e^{-\sum_{i=0}^{n}(\lambda_i+b_is)\omega_i} ds d\omega_0\cdots d\omega_n$$

$$= \frac{\prod_{i=0}^{n}\Gamma(a_i)}{\Gamma(q)} \int_0^\infty s^{q-1} \prod_{i=0}^{n} (\lambda_i + b_i s)^{-a_i} ds ,$$

因此,只要 $\sum_{i=0}^{n} a_i > q$,上述积分值是有限的,所以 K^{-1}是有限的,特别地,当 $\lambda_i=\lambda, b_i=b$ 对所有 i 成立时,就得

$$K^{-1} = \frac{\prod_{i=0}^{n}\Gamma(a_i)}{\Gamma(q)} \frac{1}{\lambda^{a-q}b^q} \frac{\Gamma(a-q)\Gamma(q)}{\Gamma(a)} ,$$

其中

$$a = \sum_{i=0}^{n} a_i ,$$

因此

$$K = \frac{\prod_{i=0}^{n}\Gamma(a_i)}{\Gamma(a)} \frac{\Gamma(a-q)}{\lambda^{a-q}b^q} .$$

从(4.8)式通过两种途径可以诱导出分布为

$$K\left(\prod_{i=0}^{n} x_i^{a_i-1}\right)\left(\sum_{i=0}^{n} b_i x_i\right)^{-q} \int_0^\infty t^{a-q-1} e^{-t\sum_{i=0}^{n} x_i\lambda_i} dt$$

$$= K\Gamma(a-q)\left(\prod_{i=0}^{n} x_i^{a_i-1}\right)\left(\sum_{i=0}^{n} b_i x_i\right)^{-q}\left(\sum_{i=0}^{n} \lambda_i x_i\right)^{-(a-q)} , \tag{4.9}$$

因此，这一类成分分布也与广义的狄氏分布有关，上式中 $a = \sum_{i=0}^{n} a_i, x_0 = 1 - \sum_{i=1}^{n} x_i$.

用条件分布的方法可诱导出另一类分布是

$$c\left(\prod_{i=0}^{n} x_i^{a_i-1}\right)\frac{e^{-\sum_{i=0}^{n}\lambda_i x_i}}{\left(\sum_{i=0}^{n} b_i x_i\right)^q}\quad ,c \text{ 是常数},$$

$$x_0 = 1 - \sum_{i=1}^{n} x_i,\quad x_i > 0,\quad i = 0,1,\cdots,n. \tag{4.10}$$

从上面的各个例子可以看出成分分布的分布族是可以有多种形式的，它并不是只限于某几类，至于哪一种分布相当于一般情况下的正态分布，起着基本而重要的作用，这还是一个值得探索的问题.

对于加法逻辑正态分布，用 (t, x) 联合分布对 t 积分求边缘分布，这已熟悉了，它是由基向量 ω 遵从对数正态诱导得来的，我们用条件密度的方法，可以得到形如下式的分布密度

$$c\left(\prod_{i=0}^{n} x_i^{-1}\right)e^{-\frac{1}{2}\sum_{i,j=0}^{n}(\ln x_i-\mu_i)(\ln x_j-\mu_j)\sigma^{ij}}$$

或

$$c\left(\prod_{i=0}^{n} x_i^{-1}\right)e^{-\frac{1}{2}(\ln x-\mu)'\sum^{-1}(\ln x-\mu)},$$

其中

$$x_0 = 1 - \sum_{i=1}^{n} x_i,\quad x_i > 0,\quad i = 0,1,\cdots,n,$$

c 是一正则化常数. 这一类分布中求出 c 的表达式比较困难.

这些例子足以说明用这两种方法可以诱导出不少成分向量的分布类型，这一讨论就到此为止.

在结束本节之前，我们要总结成分分布中常用的计算多重积分的方法，从数学分析或从黎曼积分的观点来看，一些公式的获得

是不容易，而且往往要求被积函数在连续性上有好的性质，然而从概率论或者测度论的观点看来，不需要连续性的要求，只要可测、可积(勒贝格)就可以了，而且积分公式的证明非常简便.我们就以在本书中经常用的公式为例给以说明.

著名的喀塔朗公式是：

$$\int\cdots\int_{m\leqslant g(x_1,\cdots,x_n)\leqslant M} f(x_1,\cdots,x_n)\varphi(g(x_1,\cdots,x_n))\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n$$

$$=(S)\int_m^M\varphi(u)\mathrm{d}\psi(u)=(R)\int_m^M\varphi(u)\frac{\mathrm{d}\psi(u)}{\mathrm{d}u}\mathrm{d}u, \tag{4.11}$$

式中 $(S)\int$ 表示斯蒂阶斯积分，$(R)\int$ 表示黎曼积分，$\psi(u)=\int\cdots\int_{m\leqslant g(x_1,\cdots,x_n)\leqslant M} f(x_1,\cdots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n$.

从概率论的观点来看，(4.11)的证明和意义是十分明显的.首先我们对 $f(x_1,\cdots,x_n)$可以假定为非负勒贝格可积就够了，因为任一可积的 f 一定可以写成$f=f^+-f^-$，f^+和 f^-分别是非负可积的.由于 $f(x_1,\cdots,x_n)$非负可积，其积分为有限值，于是可以无妨假定 $f(x_1,\cdots,x_n)$是随机向量$(\xi_1,\cdots,\xi_n)$的联合分布密度，假定 g,φ 都是波来儿可测函数，并且使得 $E\varphi(g(\xi_1,\cdots,\xi_n))$有限，于是 $E\varphi(g(\xi_1,\cdots,\xi_n))$可以用两种写法来表示，将 $\varphi(g(\xi_1,\cdots,\xi_n))$看成是$(\xi_1,\cdots,\xi_n)$的函数，则

$$E\varphi(g(\xi_1,\cdots,\xi_n))=\int\cdots\int f(x_1,\cdots,x_n)\times\varphi(g(x_1,\cdots,x_n))\mathrm{d}x_1\cdots\mathrm{d}x_n,$$

如果令 $\eta=g(\xi_1,\cdots,\xi_n)$，则 η 的分布函数

$$\begin{aligned}F_\eta(t)&=P(\eta\leqslant t)\\&=\int\cdots\int_{g(x_1,\cdots,x_n)\leqslant t} f(x_1,\cdots,x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n\\&=\psi(t)+F_\eta(m).\end{aligned}$$

将 $\varphi(g(\xi_1,\cdots,\xi_n))$看成 η 的函数$\varphi(\eta)$，则有

$$E\varphi(\eta)=\int\varphi(u)\mathrm{d}F_\eta(u)$$
$$=\int\varphi(u)\mathrm{d}\psi(u)$$
$$=\int\varphi(u)\frac{\mathrm{d}\psi(u)}{\mathrm{d}u}\mathrm{d}u.$$

因此对任一非负波来儿可测函数 φ 成立下式:

$$\int\cdots\int f(x_1,\cdots,x_n)\varphi(g(x_1,\cdots,x_n))\mathrm{d}x_1\cdots\mathrm{d}x_n$$
$$=\int\varphi(u)\mathrm{d}\psi(u)$$
$$=\int\varphi(u)\frac{\mathrm{d}\psi(u)}{\mathrm{d}u}\mathrm{d}u. \tag{4.12}$$

从上式不难立即得到(4.11).从概率论的观点来看,若$(\xi_1,\cdots,\xi_n)$的联合密度是$f(x_1,\cdots,x_n)$,而$\eta=g(\xi_1,\cdots,\xi_n)$的分布密度是已知的,于是求$\varphi(\eta)$的概率就是一个单积分,不论$\eta$与$(\xi_1,\cdots,\xi_n)$是否独立.这正是(4.12)式左端和右端的内容.下面我们用一些例子来说明(4.12)式的作用.

若将(4.12)理解为多重积分能用一个右端形式的单积分表示,并且$\varphi(\cdot)$是任一非负可积的函数都成立,我们就知道右端相应于$\mathrm{d}\psi(u)/\mathrm{d}u$这一函数它就是$\eta$的分布密度.所以这个等式可以有不同的作用,看是如何去用.

例 4.4 设 ξ_i 相互独立,ξ_i 各自遵从伽马分布,

$$\xi_i\sim\frac{1}{\Gamma(a_i)}x_i^{a_i-1}\mathrm{e}^{-x_i},\quad i=1,2,\cdots,n,$$

于是从概率论已知 $\eta=\sum\limits_{i=1}^n\xi_i$ 遵从伽马分布,密度是

$$\frac{1}{\Gamma(a)}u^{a-1}\mathrm{e}^{-u},\quad u>0,\quad a=\sum_{i=1}^n a_i.$$

因此从(4.12)知道,对非负可积的 $\varphi(\cdot)$,就有等式

$$\int\limits_{\substack{x_i>0\\ i=1,2,\cdots,n}}\left(\prod_{i=1}^n\frac{x_i^{a_i-1}}{\Gamma(a_i)}\right)\mathrm{e}^{-\sum\limits_{i=1}^n x_i}\varphi\left(\sum_{i=1}^n x_i\right)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

$$=\frac{1}{\Gamma(a)}\int_0^{\infty}\varphi(u)u^{a-1}\mathrm{e}^{-u}\mathrm{d}u \qquad (4.13)$$

这正是我们前面已用过多次的公式.

从这个例子看,似乎独立性起了重要的作用,其实并不是这样,我们从下一个例子看出,独立性并不是重要的.

例 4.5 设$(\xi_1,\cdots,\xi_n)$的联合密度是

$$\left(\prod_{i=1}^{n}x_i^{a_i-1}\right)q\left(\sum_{i=1}^{n}x_i\right),\quad x_i>0,\quad i=1,2,\cdots,n,$$

因此 $\xi_1,\cdots,\xi_n$ 不一定是独立的.对任一非负的可积函数 $\varphi(\cdot)$,从(4.13)就得(把这里的 $\varphi(\cdot)q(\cdot)$看成(4.13)中的 $\varphi(\cdot)$),

$$\int_{x_i>0,i=1,2,\cdots,n}\prod_{i=1}^{n}x_i^{a_i-1}q\left(\sum_{i=1}^{n}x_i\right)\varphi\left(\sum_{i=1}^{n}x_i\right)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

$$=\frac{\prod_{i=1}^{n}\Gamma(a_i)}{\Gamma(a)}\int_0^{\infty}u^{a-1}\varphi(u)q(u)\mathrm{d}u,\quad a=\sum_{i=1}^{n}a_i,$$

因此 $\eta=\sum_{i=1}^{n}\xi_i$ 的密度就是 $\frac{\prod_{i=1}^{n}\Gamma(a_i)}{\Gamma(a)}u^{a-1}q(u)$.这样用一下公式直接就求得 $\eta=\sum_{i=1}^{n}\xi_i$ 的分布,这里不要求 $\xi_1,\cdots,\xi_n$ 相互独立.

不难将上述例子形式再复杂化,考虑 $q(\sum x_i)$ 是 $q\left(\sum_{i=1}^{n}b_ix_i^{q_i}\right)$ 的形式,从而导出一些多重积分的公式,这些已不是本书的内容了.

§5. 与方向性数据、球分布的关系

现在从另一个角度来看单形上的分布.在多元统计分析中,球对称分布(简称为球分布)是一类特殊的分布,方向性数据的统计分析是一个特殊的分枝,后者涉及到在单位球球面上的分布,然而从下面的讨论中可以看出,它们与单形上的分布有密切的联系.

$n+1$ 维随机向量 $u=(u_0,u_1,\cdots,u_n)'$ 的分布称为球分布，如果对任一给定的正交方阵 Γ 有 Γu 与 u 的分布相同. 由于在正交变换群下，最大不变量是向量长度，正交变换相应的雅可比行列式绝对值是 1，因此就知道球分布如有密度，它的密度函数一定是 $f(u'u)$ 的形式. 因此我们可以用简单的符号来表示，即

$$\underset{(n+1)\times 1}{u} \sim f(u'u). \tag{5.1}$$

如果令 u 的函数：$\omega_i=u_i^2,\quad i=0,1,\cdots,n$ 作为研究的对象，则 $\omega=(\omega_0,\omega_1,\cdots,\omega_n)'$ 是正随机向量，于是从球分布(5.1)可以导出 ω 在 R_{n+1}^{+} 上的分布. 现在来看由(5.1)导出的 ω 的分布有什么特性.

引理 5.1 R_{n+1} 上的可测函数 $f(u'u)$ 若使下式左端有意义，则下述等式(5.2)就成立.

$$\int_{R_{n+1}} f(u'u)\mathrm{d}u = \int_{R_{n+1}^{+}} \left(\prod_{i=0}^{n}\omega_i^{-\frac{1}{2}}\right) f\left(\sum_{i=0}^{n}\omega_i\right)\mathrm{d}\omega_0\mathrm{d}\omega_1\cdots\mathrm{d}\omega_n$$

$$= \frac{\prod^{\frac{n+1}{2}}}{\Gamma\left(\frac{n+1}{2}\right)}\int_0^{\infty} t^{\frac{n+1}{2}-1} f(t)\mathrm{d}t. \tag{5.2}$$

证明 注意 $f(u'u)$ 的值与 u_i 的正负号无关，于是

$$\int_{R_{n+1}} f(u'u)\mathrm{d}u = 2^{n+1}\int_{R_{n+1}^{+}} f(u'u)\mathrm{d}u$$

在 R_{n+1}^{+} 上作变换：$\omega_i=u_i^2,\quad i=0,1,\cdots,n$，于是

$$\mathrm{d}\omega_i = 2u_i\mathrm{d}u_i,\quad \text{即 } \mathrm{d}u_i = \frac{1}{2}\omega_i^{-\frac{1}{2}}\mathrm{d}\omega_i,\quad i=0,1,\cdots,n$$

这样就得到

$$\int_{R_{n+1}} f(u'u)\mathrm{d}u = \int_{R_{n+1}^{+}} \left(\prod_{i=0}^{n}\omega_i^{-\frac{1}{2}}\right) f\left(\sum_{i=0}^{n}\omega_i\right)\mathrm{d}\omega_0\mathrm{d}\omega_1\cdots\mathrm{d}\omega_n$$

(5.2)的第二个等式是引理 3.1 的结论，这样我们就证明了(5.2).

(5.2)告诉我们，从球分布 $f(u'u)$ 可以导出形如

$(\prod_{i=0}^{n}\omega_i^{-\frac{1}{2}})f(\sum_{i=0}^{n}\omega_i)$ 的基向量的分布，它相应的总量 t 与成分向量 x 一定相互独立，并且成分向量 x 遵从的是一个特殊的狄氏分布.

如果反过来考虑，从 R_{n+1}^{+}上一个基向量 ω 的分布能不能引导出一个 R_{n+1}上的球分布或其他的分布呢？

从(5.2)的证明中我们已经看到如果我们要求 u 的分布对于正交群的一个子群保持不变，那么 R_{n+1}的一个密度和 R_{n+1}^{+}的密度可以有 1－1 对应的关系，这个子群就是反射群，也即

$$G = \{\Gamma = (\gamma_{ij}): \gamma_{ii} = 1 \text{ 或 } -1,$$

$$\gamma_{ij} = 0 \text{ 只要 } i \neq j \quad i = 0,1,2,\cdots,n\}.$$

这个群相应的最大不变量就是向量各个坐标的绝对值.

若 u 的密度函数是 $f(|u_0|^2,|u_1|^2,\cdots,|u_n|^2)$，$f$ 在 R_{n+1}上有定义，则令 $\omega_i = u_i^2, i=0,1,\cdots,n, \omega=(\omega_0,\cdots,\omega_n)'$后，$\omega$ 的密度在 R_{n+1}^{+}上是 $(\prod_{i=0}^{n}\omega_i^{-\frac{1}{2}})f(\omega_0,\omega_1,\cdots,\omega_n)$；反之，若先给了基向量 ω 在 R_{n+1}^{+}上的密度是函数 $p(\omega_0,\omega_1,\cdots,\omega_n)$，则令 $u_i^2=\omega_i$，$i=0,1,\cdots,n$ 后，$u=(u_0,u_1,\cdots,u_n)'$具有反射不变的密度 $(\prod_{i=0}^{n}|u_i|)p(u_0^2,u_1^2,\cdots,u_n^2)$. 当 u 的密度$f(|u_0|^2,|u_1|^2,\cdots,|u_n|^2)$只是 $\sum_{i=0}^{n}u_i^2$ 的函数时，u 就是球分布. 这样我们可以较清楚地看到 u 的分布与基向量ω 的分布之间的联系.

如再进一步考虑. 随机向量 u 的分布，是否具有反射不变的性质与 u 的方向部分有密切的关系. 实际上 u 总是可以分解为两部分：长度及方向，即

$$r_u = (\sum_{i=0}^{n}u_i^2)^{\frac{1}{2}},$$

$$\gamma_i(u) = u_i/r_u, \quad i = 0,1,\cdots,n,$$

$$\gamma_u = \begin{pmatrix} \gamma_0(u) \\ \vdots \\ \gamma_n(u) \end{pmatrix},$$

$$u = r_u \gamma_u,$$

r_u 是 u 的长度部分，γ_u 是 u 的方向部分．如果令 $\omega_i = u_i^2, i=0,1,2,\cdots,n$，于是 $\omega=(\omega_0,\omega_1,\cdots,\omega_n)'$，此时用 r_u 与 γ_u 表示 ω，就得

$$\omega_i = [\gamma_i(u)]^2 r_u^2, \quad i = 0,1,2,\cdots,n.$$

而 ω 相应的总量 t 与成分 $x=(x_0,\cdots,x_n)'$ 与 r_u，γ_u 的关系是

$$t = \sum_{i=0}^{n} \omega_i = r_u^2,$$

$$x_i = (\gamma_i(u))^2, \quad i = 0,1,\cdots,n.$$

这就明确地显示了成分与方向向量之间的密切关系．给定 R_{n+1}上的一个分布，我们可以从此分布的方向向量 γ_u 的分布诱导出一个单形上的成分分布；反之，给定了一个 R_{n+1}中单形上的成分分布，总可以由它扩展成一个 R_{n+1}上方向向量的分布，这个分布具有反射不变的特性．从这个角度来说，成分分布可以看成是方向向量分布一个特殊的部分．

从上面的分析我们还容易看出，随机向量 u 分解为长度 r_u 与方向 γ_u，这与基向量 ω 分解为总量 t 与成分向量 x 是完全相对应的．而球分布在 R_{n+1}中占有重要的地位，它相应的是 r_u 与 γ_u 相互独立，而且 γ_u 是球面上的均匀分布．在基向量 ω 的分布中，与之相应的是总量 t 与成分相互独立，成分服从狄氏分布．从这一比较中也可以看出狄氏分布的重要．

在 R_{n+1}中，单位球球面上的点用坐标来表示，坐标记为$(u_0,u_1,\cdots,u_n)'$，则有

$$\sum_{i=0}^{n} u_i^2 = 1,$$

令

$$x_i = u_i^2, \quad i = 0,1,2,\cdots,n, \text{则} \sum_{i=0}^{n} x_i = 1,$$

且

$$x_i \geqslant 0, i = 0,1,\cdots,n.$$

可见球面上的点可以与单形上的点发生对应;反之单形上的点也可与球面上的点对应.现在利用这种对应关系来看一些成分向量的分布.

例 5.1 球面上的均匀分布与单形上的均匀分布.

从球分布的性质我们知道标准正态分布

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{n+1} e^{-\frac{1}{2}\sum_{i=0}^{n} y_i^2}$$

是 R_{n+1} 中的球分布,它的长度部分与方向部分相互独立.长度部分用 r 表示,则 $r^2 \sim \chi^2(n+1)$,方向部分用$(u_0,u_1,\cdots,u_n)$表示,则$(u_1,\cdots,u_n)$的联合分布密度是

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\pi^{\frac{n}{2}}}\left(1-\sum_{i=1}^{n} u_i^2\right)^{-\frac{1}{2}}, \quad \sum_{i=1}^{n} u_i^2 < 1 \tag{5.3}$$

而且知道$(u_0,u_1,\cdots,u_n)$是球面上的均匀分布.因此(5.3)就给出了球面上均匀分布的密度的表示式.利用 $x_i = u_i^2, \mathrm{d}x_i = 2u_i\mathrm{d}u_i, i = 1,2,\cdots,n$,从(5.3)式可以导出单形上$(x_0,x_1,\cdots,x_n)$的联合分布密度,而且它也就是单形上的均匀分布,计算一下雅可比行列式就得到

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\pi^{\frac{n}{2}}}\left(\prod_{i=1}^{n} x_i^{-\frac{1}{2}}\right)\left(1-\sum_{i=1}^{n} x_i\right)^{-\frac{1}{2}}, \quad x_i > 0, \sum_{i=1}^{n} x_i < 1,$$

即为

$$\Gamma\left(\frac{n+1}{2}\right)\pi^{-\frac{n+1}{2}}\prod_{i=0}^{n} x_i^{-\frac{1}{2}}, \quad x_i > 0, \sum_{i=1}^{n} x_i < 1, \sum_{i=1}^{n} x_i = 1 \tag{5.4}$$

这是狄氏分布的一个特殊情形.

从这个例子可以看出,球面上的分布可以用低一维的单位球

上的分布来描述,因此单形 S_{n+1}上的分布可以用低一维的 D_n 上的分布来描述. 只要对单位球上的分布有所讨论,就相当于对 D_n 上的分布有所讨论,也就是对单形 S_{n+1}上的分布有所讨论. 因此下面的积分公式就可以起相当重要的作用.

$$\int_{R_n}\Big(\prod_{i=1}^{n}|u_i|^{2a_i-1}\Big)f(u'u)\mathrm{d}\underset{n\times1}{u}=\int_{R_n^+}\Big(\prod_{i=1}^{n}x_i^{a_i-1}\Big)f\Big(\sum_{i=1}^{n}x_i\Big)\mathrm{d}\underset{n\times1}{x}$$

$$=\frac{\prod_{i=1}^{n}\Gamma(a_i)}{\Gamma\Big(\sum_{i=1}^{n}a_i\Big)}\int_0^{\infty}t^{\sum_{i=1}^{n}a_i-1}f(t)\mathrm{d}t. \tag{5.5}$$

这一公式的证明只需要将上式左端积分利用对称性,可表示为

$$2^n\int_{R_n^+}\Big(\prod_{i=1}^{n}u_i^{2a_i-1}\Big)f(u'u)\mathrm{d}u,$$

再作变换

$$x_i = u_i^2,\quad i = 1,2,\cdots,n,\quad \mathrm{d}x_i = 2u_i\mathrm{d}u_i,$$

得到

$$2^n\int_{R_n^+}\Big(\prod_{i=1}^{n}u_i^{2a_i-1}\Big)f(u'u)\mathrm{d}u = \int_{R_n^+}\Big(\prod_{i=1}^{n}x_i^{a_i-1}\Big)f\Big(\sum_{i=1}^{n}x_i\Big)\mathrm{d}\underset{n\times1}{x}$$

上式右端是已知的,它的值就是(5.5)的右端.

有了这个(5.5)式,就可以将单位球上的一个分布密度与单形上的分布密度建立对应的关系.

在 R_n 中单位球 $\sum_{i=1}^{n}u_i^2<1$ 上的分布密度,一定可以写成

$$\Big(\prod_{i=1}^{n}|u_i|^{2a_i-1}\Big)I(u'u<1)f(u_1,\cdots,u_n)$$

的形式,其中

$$I(u'u<1)=\begin{cases}1, & \text{当 } u'u<1,\\ 0, & \text{其他}.\end{cases}$$

因此只要 $f(u_1,\cdots,u_n)$是 $u'u$ 的函数,它有 $h(u'u)$的表达式,则

从(5.5)式就引导出一个在 D_n 上的分布密度

$$\left(\prod_{i=1}^{n} x_i^{a_i-1}\right)f\left(\sum_{i=1}^{n} x_i\right),\quad x_i>0,\quad \sum_{i=1}^{n} x_i<1.$$

写成与 S_{n+1} 有关的表示式,就是

$$\left(\prod_{i=1}^{n} x_i^{a_i-1}\right)f(1-x_0),\quad x_i>0,\quad i=0,1,\cdots,n,\quad \sum_{i=0}^{n} x_i=1,$$

从这里不难引出一些别的成分分布的密度.

90 年代以来,多元统计分析中出现一类分布称为径可分(radially decomposable)分布.这一类分布的特点是与 α 对称分布族有关,与成分向量的分布也有关.现给出它的定义并讨论它与成分向量的联系.

定义 5.1[4] 随机向量 $\underset{n\times 1}{z}$ 称为是具有径可分分布的随机向量,如果存在正的随机变量 R,另一个与 R 独立的随机向量 $n\times 1$ 的 w,使得

$$z \overset{d}{=} Rw, \tag{5.6}$$

即 z 与 Rw 具有相同的分布.此时称 R 为 z 的尺度(scalar),w 称为基本向量(base vector).

回忆一下我们在前面讨论过的形状与大小,这两者又有区别又有联系,有关这一方面的讨论留给读者去考虑.我们讨论径可分的 z 有什么特殊的性质.

引理 5.2[4] 设 $t(z)$是 z 的函数,且

$$t(az)=t(z) \tag{5.7}$$

对一切 $a>0$ 成立,则 $t(z) \overset{d}{=} t(\omega)$.

这条引理告诉我们,只要 $t(z)$是一个齐 0 次的函数,或者说它对尺度变换(即自变量乘一个正的常数)不变,那么它的分布就只与 w 有关.现在来给出证明.

证明 考虑特征函数

$$\begin{aligned}\varphi_{t(z)}(\theta) &= E\mathrm{e}^{i\theta' t(z)} = E(E\{\mathrm{e}^{i\theta' t(R\omega)}\mid R\})\\ &= E(E\{\mathrm{e}^{i\theta' t(\omega)}\mid R\})\end{aligned}$$

$$
\begin{aligned}
&= Ee^{i\theta' t(\omega)} \\
&= \varphi_{t(\omega)}(\theta).
\end{aligned}
$$

这一性质引出很多有用的结论，有关的内容可以参阅文献[4].

注意到我们一直讨论的基向量(不是基本向量)w 分解为总量 t 与成分向量 x 时，总量 t 与成分独立，就表示 w 是一个径可分的随机向量，我们对比一下就清楚了：

$$
\begin{aligned}
\text{基向量 } w &\longleftrightarrow \text{随机向量 } z \\
w = tx &\longleftrightarrow z \overset{d}{=} Rw, \quad R > 0 \\
t \text{ 与 } x \text{ 独立} &\longleftrightarrow R \text{ 与 } w \text{ 独立},
\end{aligned}
$$

差别是两条：

(i)基向量 w 要求它的分量都是正的随机变量，而对径可分的 z 并不要求；

(ii)$w = tx$ 是一个真实的等式，而 $z \overset{d}{=} Rw$ 是一个在分布意义上的等式，后者的要求就更弱一些.

从一定的意义上看，径可分的概念是推广了成分与总量，形状与大小，引导到一个更一般的情形.

习　题　二

1. 由伽马函数的定义

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}\mathrm{d}t, \quad a > 0,$$

直接证明

$$\int_0^\infty t^{ab-1}e^{-\beta t^b}\mathrm{d}t = \frac{\Gamma(a)}{b\beta^a}, \quad a > 0, \quad b > 0, \quad \beta > 0.$$

如果 $b<0$，上述等式应如何修改？对 $b\neq0, a>0, \beta>0$，能否有一个统一的公式.

2. 如果基向量 ω 的大小 $G(\omega)$ 与形状 $z_G(\omega)$ 相互独立，则对任一大小 $G_*(\omega)$，$G(\omega)$ 一定与 $z_{G_*}(\omega)$ 相互独立. 进一步考虑 $z_{G_*}(\omega)$ 与 $z_G(\omega)$ 分布密度的关系式.

3. 对于二分割的逻辑正态分布共有几种？写出相应的密度. 进一步讨

论三分割的分布密度有多少种不同的，能否给出一般 K 分割的不同密度的个数(注意每个分割的子成分可以各自选用加法逻辑正态或乘法逻辑正态，此外还有子成分总量形成的向量 t，它也有两种选择).

4. 证明等式(3.5).

5. 若基向量的分量 ω_i 相互独立，且

$$\omega_i \sim \frac{\lambda_i^{a_i}}{\Gamma(a_i)}\omega_i^{-(a_i+1)}\mathrm{e}^{-\frac{\lambda_i}{\omega_i}}, \quad i = 0,1,\cdots,n.$$

求它相应的成分向量的分布密度.

6. 若基向量 $\omega=(\omega_0,\omega_1,\cdots,\omega_n)'$ 的联合密度是

$$\frac{1}{C(\boldsymbol{a};f)}\Big(\prod_{i=0}^{n}\omega_i^{a_i-1}\Big)f\Big(\sum_{i=0}^{n}\omega_i\Big), \quad \omega_i>0, \quad i=0,1,\cdots,n,$$

其中 $C(\boldsymbol{a};f)$ 为一常数，$\boldsymbol{a}=(a_0,a_1,\cdots,a_n)'$.

(i) 求出 $C(\boldsymbol{a};f)$ 的表达式.

(ii)若 $a = \sum_{i=0}^{n} a_i$ 是(0,1)区间内的值，

$$f\Big(\sum_{i=0}^{n}\omega_i\Big) = \Big(1+\sum_{i=0}^{n}\omega_i\Big)^{-1}.$$

求证：$C(\boldsymbol{a};f) = \dfrac{\Gamma(a)}{\prod_{i=0}^{n}\Gamma(a_i)}\dfrac{\sin a\pi}{\pi}$.

(iii) 若 $a = \sum_{i=0}^{n} a_i$，且

$$f(t) = t^{-(a-1)}\mathrm{e}^{-t^2-(1/t^2)}.$$

求证：$C(\boldsymbol{a};f) = \dfrac{\Gamma(a)}{\prod_{i=0}^{n}\Gamma(a_i)}\dfrac{2\mathrm{e}^2}{\sqrt{\pi}}$.

从这几个不同的 f 选择中，你可以看出，成分向量服从狄氏分布时，基向量的联合分布可以相当复杂而广泛.

((ii),(iii)两题的积分可参看 Г. М. 菲赫金哥尔茨的微积分学教程第十四章)

7. 证明：当基向量 ω 的分布是(4.4)式函数时，成分向量 x 对总量 t 的条件密度是(4.6).

8. 若 $\ln x$ 与 $\ln y$ 是正态分布，且 $\frac{x}{y}$ 与 x 独立，$\frac{x}{y}$ 又与 y 独立，则 $\frac{x}{y}$ 概率为 1 是一个常数.(给出与定理 1.2 不同的证明，提示：证明

$$\mathrm{Var}(\ln x - \ln y) = 0)$$

9. 若基向量是相互独立的分量，每个分量都是广义反正态分布，试用两种方法导出相应的成分分布密度.

10. 用本章的(4.12)导出(4.11).

11. 对$(\ln x_1, \ln x_2, \ln x_3)$这个向量，不存在对数比不相关，但存在对数对比不相关；对$(\ln x_1, \ln x_2, \ln x_3, \ln x_4)$这两者关系又是怎样呢？

12. 讨论径可分随机向量的形状与大小的关系. 利用这个结果可以导出一些什么结论(将它用于径可分向量：如标准正态分布的情形).

参 考 文 献

[1] Aitchison, J. (1986), The Statistical Analysis of Compositional Data, Chapman & Hall.

[2] Grow, E. L. and Shimizu, K. (1988), Lognormal Distributions, Theory and applications. Marcel Dekker, INC. New York and Basel.

[3] 菲赫金哥尔茨，Γ. M. (1954)，微积分学教程，高等教育出版社.

[4] Ng, K. W. and Fraser, D. A. S. (1994), Inference for linear models with radially decomposable error, Multivariate Analysis and Its Applications, IMS Lecture Notes-Monograph Series, Vol. 24.

第三章　逻辑正态分布的统计分析

§1. 估　　计

本章限于成分向量 x 服从逻辑正态分布的情形，以下不再逐一声明.

为了用 n 表示样本容量的大小，总假定成分向量是 $p+1$ 维，$x=(x_0,x_1,\cdots,x_p)'$，x 的样本用矩阵 X 来表示，X 的每一行是一个样本，即

$$\underset{n\times(p+1)}{X}=\begin{pmatrix}x'_{(1)}\\ \vdots\\ x'_{(n)}\end{pmatrix}=(x_{\alpha i})\quad \alpha=1,2,\cdots,n,\quad i=0,1,\cdots,p.\tag{1.1}$$

于是

$$y=(-\mathbb{1}\quad I_p)\ln x=(-\mathbb{1}\quad I_p)\begin{pmatrix}\ln x_0\\ \vdots\\ \ln x_p\end{pmatrix}$$

遵从正态分布 $N(\mu,\Sigma)$. y 相应的样本用矩阵 Y 表示，也即

$$\underset{n\times p}{Y}\triangleq(\ln X)\begin{pmatrix}-\mathbb{1}'\\ I_p\end{pmatrix}=(\ln x_{\alpha i})\begin{pmatrix}-\mathbb{1}'\\ I_p\end{pmatrix}\triangleq(\ln X)F,\tag{1.2}$$

它是由 X 加工生成的.

现在来看成分向量的估计问题与一般多元统计分析的估计有什么不同，应该怎样处理.

描述成分向量 x 的参数，自然是它的期望值向量，它的协方差矩阵，这两组参数是重要的. 我们用 θ 表示 Ex，$\mathscr{K}$表示 x 的协方差矩阵，即

$$\underset{(p+1)\times 1}{\theta} = \begin{pmatrix} Ex_0 \\ \vdots \\ Ex_p \end{pmatrix}, \underset{(p+1)\times(p+1)}{\mathscr{H}} = (\eta_{ij}) = (E(x_i - Ex_i)(x_j - Ex_j)),$$

$$i,j = 0,1,\cdots,p.$$

现在的问题是,如何求 θ 与 $\mathscr{H}$ 的估计量 $\hat{\theta}$ 与 $\hat{\mathscr{H}}$.

注意到 $\sum_{i=0}^{p} x_i = 1$ 总是成立的,因此自然有

$$\begin{cases} 1 = E\left(\sum_{i=0}^{p} x_i\right) = \sum_{i=0}^{p} Ex_i = \mathbb{1}'\theta, \\ 0 = E\,\mathbb{1}'(x - Ex)(x - Ex)'\,\mathbb{1} \\ \quad = \mathbb{1}'E(x - Ex)(x - Ex)'\,\mathbb{1} = \mathbb{1}'\mathscr{H}\mathbb{1}. \end{cases} \tag{1.3}$$

上式由于 $\mathscr{H}$ 是一非负定矩阵,它就等价于 $\mathscr{H}\mathbb{1}=0$,即 $\mathbb{1}$ 是 $\mathscr{H}$ 的特征值为 0 相应的特征向量.这表明参数 θ 与 $\mathscr{H}$ 是受约束的.由第一章习题 6 的结论知道一定存在正定阵 Ω,使得

$$\mathscr{H} = \Omega^{-1} - \Omega^{-1}\mathbb{1}\mathbb{1}'\Omega^{-1}/(\mathbb{1}'\Omega^{-1}\mathbb{1}).$$

因此我们将 θ 与 $\mathscr{H}$ 的约束条件(1.3)可以写成

$$\begin{cases} \theta'\mathbb{1} = 1, \quad \theta_i \geqslant 0, \quad i = 0,1,2,\cdots,p, \\ \mathscr{H} = \Omega^{-1} - \Omega^{-1}\mathbb{1}\mathbb{1}'\Omega^{-1}/(\mathbb{1}'\Omega^{-1}\mathbb{1}), \quad \Omega > 0. \end{cases}$$

现在来寻求满足条件(1.3)的估计量 $\hat{\theta}$ 和 $\hat{\mathscr{H}}$.从统计的观点来看,有三种方法可以导出相应的估计.

(i)矩估计法

用样本的均值去估计总体的均值,用样本的协差阵去估计总体的协差阵,于是有

$$\begin{cases} \hat{\theta} = \frac{1}{n}X'\mathbb{1} \triangleq \bar{x}, \\ \hat{\mathscr{H}} = \frac{1}{n}\sum_{\alpha=1}^{n}(x_{(\alpha)} - \bar{x})(x_{(\alpha)} - \bar{x})' \\ \quad = \frac{1}{n}X'\left(I - \frac{1}{n}\mathbb{1}\mathbb{1}'\right)X \triangleq \frac{1}{n}L_{xx}. \end{cases} \tag{1.4}$$

现在来看(1.4)的估计是否合于(1.3)的要求. 注意到 X 的每一行都是成分向量的样本，自然有

$$X\mathbb{1} = \mathbb{1}.$$

因此$\mathbb{1}'\hat{\theta} = \mathbb{1}'\bar{x} = \frac{1}{n}\mathbb{1}'X'\mathbb{1} = \frac{1}{n}\mathbb{1}'\mathbb{1} = 1, \hat{\mathscr{H}}\mathbb{1} = \frac{1}{n}L_{xx}\mathbb{1} = 0$，$\hat{\mathscr{H}}$是非负定的. 可见(1.4)给出的估计是合于(1.3)的条件的.

(ii)利用 y 的样本来给出估计

因为 $y \sim N(\mu, \Sigma)$，因此从正态分布的性质就知道 y 的样本均值和样本协差阵是 μ 和Σ的优良估计，这就得 μ 和Σ的估计量为

$$\begin{cases} \hat{\mu} = \bar{y} = \frac{1}{n}Y'\mathbb{1} = \frac{1}{n}F'(\ln X)'\mathbb{1}, \\ \hat{\Sigma} = \frac{1}{n-1}Y'\left(I - \frac{1}{n}\mathbb{1}\mathbb{1}'\right)Y = \frac{1}{n-1}L_{yy}, \end{cases} \tag{1.5}$$

其中

$$F' = (-\mathbb{1} \quad I_p),$$

$$L_{yy} = F'(\ln X)'\left(I - \frac{1}{n}\mathbb{1}\mathbb{1}'\right)(\ln X)F \triangleq F'L_{\ln x \ln x}F.$$

问题是参数 μ，Σ与参数 θ，$\mathscr{H}$之间的关系如何表示，怎样利用这个关系式导出用 $\hat{\mu}$，$\hat{\Sigma}$去求$\hat{\theta}$ 和 $\hat{\mathscr{H}}$.

要求 μ，Σ与 θ，$\mathscr{H}$之间的关系，必然涉及到 y 与 x 的关系，y 与 x 的联系是

$$y_i = \ln x_i - \ln x_0 = \ln\frac{x_i}{x_0}, \quad i = 1,2,\cdots,p, \tag{1.6}$$

它也可以写成

$$\frac{x_i}{x_0} = e^{y_i}, \quad i = 1,2,\cdots,p. \tag{1.7}$$

这两个不同的表示法可引出两种不同的估计.

用 $\ln(1+x) \doteq x$ 的近似式，从(1.6)就可以得

$$y_i \doteq x_i - x_0.$$

因为

$$y_i = \ln x_i - \ln x_0 = \ln(1 + (x_i - 1)) - \ln(1 + (x_0 - 1))$$
$$\doteq x_i - 1 - x_0 + 1 \doteq x_i - x_0,$$

因此

$$\mu_i = Ey_i \doteq Ex_i - Ex_0 = \theta_i - \theta_0, \quad i = 1,2,\cdots,p,$$
$$\sum_{i=1}^{p} \mu_i = \sum_{i=1}^{p} \theta_i - p\theta_0 = 1 - (p+1)\theta.$$

因而就有

$$\begin{cases} \theta_i = \mu_i + \theta_0, & i = 1,2,\cdots,p, \\ \theta_0 = \left(1 - \sum_{i=1}^{p} \mu_i\right) \Big/ (p+1). \end{cases} \tag{1.8}$$

将 $\hat{\mu}$ 用 $\bar{y}$ 代入上式，可得 $\hat{\theta}_i$ 及 $\hat{\theta}_0$. 如此求得的 $\hat{\theta}$ 能满足 $1'\hat{\theta} = 1$，但 $\hat{\theta}_i$ 或 $\hat{\theta}_0$ 可能会出现负值，这是不合要求的.

当然还可以将 $\ln x_i = \ln(1 + (x_i - 1))$ 多展开几项来求关系式，然而这样的讨论就很繁琐，这里就不展开了.

用(1.7)式，利用 y 是正态的性质，知道

$$E\frac{x_i}{x_0} = Ee^{y_i} = e^{\mu_i + \frac{1}{2}\sigma_{ii}}, \quad i = 1,2,\cdots,p,$$

将上式两端对 i 求和得

$$E\frac{1 - x_0}{x_0} = \sum_{i=1}^{p} e^{\mu_i + \frac{1}{2}\sigma_{ii}},$$

于是有

$$E\frac{1}{x_0} = 1 + \sum_{i=1}^{p} e^{\mu_i + \frac{1}{2}\sigma_{ii}}.$$

利用算术平均、调和平均不等式，由于 $x_0 > 0$，得

$$Ex_0 \geqslant (Ex_0^{-1})^{-1} = \left(1 + \sum_{i=1}^{p} e^{\mu_i + \frac{1}{2}\sigma_{ii}}\right)^{-1}.$$

将上述不等号看成等号，就可以引出 θ_i 的估计

$$\begin{cases} \hat{\theta}_0 = \left(1 + \sum_{i=1}^{p} e^{\bar{y}_i + \frac{1}{2}\hat{\sigma}_{ii}}\right)^{-1}, \\ \hat{\theta}_i = e^{\bar{y}_i + \frac{1}{2}\hat{\sigma}_{ii}} \Big/ \left(1 + \sum_{j=1}^{p} e^{\bar{y}_j + \frac{1}{2}\hat{\sigma}_{jj}}\right), \quad i = 1,2,\cdots,p . \end{cases} \tag{1.9}$$

(1.9)式的估计不会给出不合理的非负估计，而且它与 y 的样本均值 $\bar{y}$ 及样本协差阵的主对角元素 $\hat{\sigma}_{ii}$，即

$$\hat{\Sigma} = \frac{1}{n-1} L_{yy} = (\hat{\sigma}_{ij})$$

中的主对角元有关.(1.9)式合乎约束条件 $\mathbb{1}'\hat{\theta} = 1$，因此它是利用了 y 是正态分布的这一特性导出的.然而也不难看出，这一估计还不是很理想的，它也是一种近似的估计.

从对数正态分布的性质知道(见第一章(4.6)式)

$$\mathrm{Cov}\left(\frac{x_i}{x_0}, \frac{x_j}{x_0}\right) = \left(\exp\left\{\mu_i + \mu_j + \frac{1}{2}(\sigma_{ii} + \sigma_{jj})\right\}\right)(e^{\sigma_{ij}} - 1) ,$$

于是从 $\hat{\mu}$ 及 $\hat{\Sigma}$ 可以给出 $\left(\frac{x_1}{x_0}, \cdots, \frac{x_p}{x_0}\right)$ 所相应的协方差矩阵的估计.但要给出 $\mathscr{K}$ 的估计就更麻烦了.从这里可以看到成分数据统计分析确有它的特殊困难.

(iii) 利用 x_i 与 y_i 之间的关系式，从大样本角度考虑给出的估计.

因为 x_i 与 y_i 之间有关系式

$$\begin{cases} x_i = e^{y_i} \Big/ \left(1 + \sum_{j=1}^{p} e^{y_j}\right) , \quad i = 1,2,\cdots,p , \\ x_0 = \left(1 + \sum_{j=1}^{p} e^{y_j}\right)^{-1}. \end{cases} \tag{1.10}$$

因此

$$Ex_i = E\left(e^{y_i}\left(1 + \sum_{j=1}^{p} e^{y_j}\right)^{-1}\right) \triangleq Ef_i(y) , \quad i = 1,2,\cdots,p ,$$
$$f_i(y) = f_i(y_1, \cdots, y_p)$$

$$= f_i(Ey_1,\cdots,Ey_p) + \sum_{j=1}^{p} \frac{\partial f_i}{\partial y_j}(y_j - Ey_j) + r \ ,$$

其中右端 r 为展式的余项,略去余项 r,对二端求期望,就得近似式

$$Ef_i(y) \doteq Ef_i(Ey_1,\cdots,Ey_p) = f_i(Ey_1,\cdots,Ey_p) \ ,$$

也即有

$$\begin{cases} \theta_i \doteq e^{\mu_i} \Big/ \left(1 + \sum_{j=1}^{p} e^{\mu_j}\right) , & i = 1,2,\cdots,p \ , \\ \theta_0 \doteq \left(1 + \sum_{j=1}^{p} e^{\mu_j}\right)^{-1}. \end{cases} \tag{1.11}$$

于是就得 θ 的估计量

$$\begin{cases} \hat{\theta}_i = e^{\bar{y}_i} \Big/ \left(1 + \sum_{j=1}^{p} e^{\bar{y}_j}\right) , & i = 1,2,\cdots,p \ , \\ \hat{\theta}_0 = \left(1 + \sum_{j=1}^{p} e^{\bar{y}_j}\right)^{-1}, \end{cases} \tag{1.12}$$

$\bar{y} = (\bar{y}_1,\cdots,\bar{y}_p)' = \frac{1}{n}Y'\mathbb{1}$. 易见(1.12)给出的估计总是非负的,而且合于$\mathbb{1}'\hat{\theta} = 1$ 的要求.

再稍加仔细看一下,将 $\bar{y}_i$ 用成分向量的样本表示一下,就有

$$\bar{y}_i = \frac{1}{n}\sum_{\alpha=1}^{n}(\ln x_{\alpha i} - \ln x_{\alpha 0}) \triangleq \ln g_i - \ln g_0, \quad i = 1,2,\cdots,p \ ,$$

其中

$$g_i = \left(\prod_{\alpha=1}^{n} x_{\alpha i}\right)^{\frac{1}{n}}, \quad i = 0,1,\cdots,p \ .$$

这样就有

$$e^{\bar{y}_i} = g_i / g_0, \quad i = 1,2,\cdots,p \ ,$$

而(1.12)就是

$$\hat{\theta}_i = g_i \Big/ \left(\sum_{j=0}^{p} g_j\right) , \quad i = 0,1,\cdots,p \ .$$

将(1.12)与(1.4)作比较,可见(1.4)是由算术平均得来的,而

(1.12)是以成分资料的几何平均为权再生成估计的.后者利用了逻辑正态分布的特性.

(1.9)式的估计在某种意义上是用了调和平均给出的.这几种不同的估计会有多大的差别,下面我们用一个数字的例子给以说明,使人有一个直观的印象.

下面是甲、乙两个医院分科治疗效果的记录,年份是1988—1990三年,其中成分变量共五个,即治疗效果:

x_1　　治愈百分比

x_2　　好转百分比

x_3　　无效百分比

x_4　　死亡百分比

x_5　　未治疗百分比

从这组数据来说明一些估计量给出估计的差异情况.详细列出原始记录就太长了,这里扼要地对原始数据给出描述,着重于估计量的数值有多大的差别.

表 1.1　两医院的治疗情况(各变量的年均值)

科别	年份	甲医院(‰)					乙医院(‰)				
		x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
内科	1988	405	441	023	047	084	289	571	047	030	063
	1989	380	494	021	047	059	262	555	046	028	109
	1990	370	493	018	039	080	252	576	025	044	103
外科	1988	801	077	012	016	094	714	129	024	013	120
	1989	785	079	011	022	103	664	114	026	022	174
	1990	805	068	008	011	108	705	111	017	012	155
传染科	1988	855	112	002	019	011	834	115	015	028	008
	1989	794	143	007	024	032	748	163	028	039	022
	1990	841	107	007	023	021	696	189	048	041	027
妇科	1988	848	084	007	006	055	836	065	004	003	092
	1989	818	108	003	005	065	809	056	005	003	128
	1990	784	162	000	002	052	783	079	005	004	128

续表 1.1

科别	年份	甲医院(‰)					乙医院(‰)				
		x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
儿科	1988	781	123	013	043	040	790	170	021	007	011
	1989	672	222	013	035	059	705	247	019	010	019
	1990	657	218	024	045	056	709	212	022	011	045
全院	1988	702	186	013	023	075	622	251	031	017	079
	1989	617	199	012	021	150	568	264	034	019	115
	1990	657	179	009	019	136	564	270	024	023	118

从数据来看,成分确实呈现一种稳定性,逐年变化总的来说不是很大.上述数据是按每年各个月份计算后,再算年度的月平均值,所以已经是加工了的资料.再将逐月的成分数据取对数,再求出均值和方差,就相当于观察数据用 $y=\ln x$ 所得的样本均值和方差.

为了便于核实,就用表 1.1 的数据作为原始数据,把甲、乙两医院的资料合并,传染科与妇科的数据是相近的,把这两科的数据也合并,这样就有 5 个变量,12 个样本,算出成分数据 x_i 的均值、标准差的数据;$y_i=\ln x_i$ 的均值、标准差的数据.结果见表 1.2,1.3(注意我们是用‰表示的).

表 1.2　x_i 的均值、标准差

变量	均值	标准差
x_1	803.83	46.51
x_2	115.25	41.70
x_3	10.92	13.85
x_4	16.42	14.51
x_5	53.41	42.42

表 1.3　$y_i=\ln x_i$ 的均值、标准差

变量	均值	标准差
y_1	6.6878	5.979×10^{-2}
y_2	4.6837	0.3791
y_3	2.0133	0.9480
y_4	2.3038	1.1219
y_5	3.6444	0.9053

它们各自的方差、协方差矩阵、相关矩阵分别为 $V(x)$，$V(y)$，$R(x)$，$R(y)$，相应的数值是

$$V(x)=\begin{bmatrix} 2162.879 & -1359.045 & -508.288 & -348.924 & 36.621 \\ & 1739.295 & 371.295 & 426.341 & -1171.568 \\ & & 191.720 & 163.674 & -213.144 \\ & & & 210.629 & -449.371 \\ & & & & 1800.811 \end{bmatrix},$$

$$R(x)=\begin{bmatrix} 1.000 & -0.701^{*} & -0.789^{**} & -0.517 & 0.19 \\ & 1.000 & 0.643^{*} & 0.704^{*} & -0.662^{*} \\ & & 1.000 & 0.814^{**} & -0.363 \\ & & & 1.000 & -0.730^{*} \\ & & & & 1.000 \end{bmatrix},$$

$$V(y)=\begin{bmatrix} 3.575\times10^{-3} & -1.35\times10^{-2} & -4.57\times10^{-2} & -2.06\times10^{-2} & 3.45\times10^{-3} \\ & 0.144 & 0.229 & 0.258 & -0.199 \\ & & 0.899 & 0.645 & -0.280 \\ & & & 1.259 & -0.816 \\ & & & & 0.820 \end{bmatrix},$$

$$R(y)=\begin{bmatrix} 1.000 & -0.598^{*} & -0.774^{**} & -0.308 & -0.064 \\ & 1.000 & 0.644^{*} & 0.607^{*} & -0.581^{*} \\ & & 1.000 & 0.648^{*} & -0.313 \\ & & & 1.000 & -0.803^{**} \\ & & & & 1.000 \end{bmatrix}.$$

* 表示 0.05 水平显著，** 表示 0.01 水平显著（H_0 是相关系数为 0）

将 $x_1,x_2,\cdots,x_5$ 中哪一个选作 x_0 去求 $y_i=\ln x_i/x_0$，显然给出的 y 均值是不一样的.因此所得的估计值就不同.另一点要注意的是妇科 90 年、甲医院的 $x_3=0$，这是观察的 0，理论上不会是 0，而且取对数时 $\ln 0=-\infty$ 就无法处理，因此将这个数据用 1‰代入，或当作缺失数据处理，用最小二乘补空.再一点可以看出 $R(x)$ 与 $R(y)$ 的显著性部分没有差异，但显著的程度有差别.从相关系数的正负号来看，也是有意思的，不少变量之间存在显著的正相关.

我们分别用 x_5 作为 x_0，x_4 作为 x_0，再用公式(1.9)和(1.12)

给出估计值,列表与 x_i 的算术平均值比较,这就是下面的表 1.4.

表 1.4　不同估计方法的估计值

	Ex_1	Ex_2	Ex_3	Ex_4	Ex_5	总和
算术平均	803.83	115.25	10.92	16.42	53.41	999.83
$\ln x_i/x_5(1.9)$	781.81	137.50	14.91	41.17	24.61	1000.00[*]
$\ln x_i/x_5(1.12)$	831.38	112.06	6.56	10.37	39.64	1000.01
$\ln x_i/x_4(1.9)$	781.00	85.43	4.48	5.05	124.04	1000.00[*]
$\ln x_i/x_4(1.12)$	832.32	110.97	6.50	10.27	39.25	999.31

[*] 这是用 1 去减全部和得到的.

表 1.4 中总和不是 1000,是由于计算数值时四舍五入引起的误差.

§2. 期望值检验

对于成分数据,假定总体分布是加法逻辑正态时,期望值检验的问题可以看成是正态总体期望值的检验问题,因此不难从正态总体的一些检验方法直接导出有关成分向量期望值检验的方法.

例如,成分向量 $\underset{(p+1)\times 1}{x}$ 的期望值是 $\theta = Ex$,它的对数变换 $y_i = \ln x_i/x_0$, $i = 1,2,\cdots,p$ 是正态分布,因此要检验两个总体相应的 x 的期望值相同,也就是两个总体相应的 y 的期望值相同,因为 y 与 x 之间是双方 1-1 的变换,问题是等价的.这样,就不用对 x 作检验,只消将 x 数据变换为 y,直接对 y 用正态分布的一些结论就可以了.很明显,对协方差矩阵的检验也是如此,这样考虑,似乎成分数据的检验问题,在加法逻辑正态分布的假定下,就很好解决了.实际情况并非如此,我们先用举例的方式列出有关用正态结论的检验方法,然后再进一步讨论为什么这样还有问题.

沿用上一节的记号,成分数据的样本矩阵 X 加工为 Y:

$n \times p$，Y 的各行独立同分布，来自 $N(\mu, \underset{p\times p}{\Sigma})$.

例 2.1 已知协方差矩阵Σ，要检验

$$H_0: \mu = \mu_0; \quad H_1: \mu \neq \mu_0.$$

此时使用统计量

$$n(\bar{y} - \mu_0)' \Sigma^{-1} (\bar{y} - \mu_0), \quad \bar{y} = \frac{1}{n} Y' \mathbb{1}, \tag{2.1}$$

H_0 成立时，它的分布是 $\chi^2(p)$——p 个自由度的 χ^2 分布.

例 2.2 未知协方差矩阵Σ，要检验

$$H_0: \mu = \mu_0; \quad H_1: \mu \neq \mu_0.$$

此时使用 T^2 统计量，

$$\begin{aligned} T^2 &= n(n-1)(\bar{y} - \mu_0)' S^{-1} (\bar{y} - \mu_0), \\ S &= Y'\left(I_n - \frac{1}{n} \mathbb{1}\mathbb{1}'\right) Y, \end{aligned} \tag{2.2}$$

当 $n > p$ 时，(2.2)给出 T^2 有：H_0 成立时，

$$\frac{n-p}{(n-1)p} T^2 \sim F(p, n-p).$$

例 2.3 多总体的均值检验.

假定 Y_i 是来自总体$N(\mu_i, \Sigma)$的样本矩阵，Y_i 的大小是$n_i \times p$，$i = 1,2,\cdots,k$，Σ未知，此时要检验

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k. \tag{2.3}$$

记

$$\bar{y}_i = \frac{1}{n_i} Y_i' \mathbb{1}, \quad S_i = Y_i'\left(I - \frac{1}{n_i} \mathbb{1}\mathbb{1}'\right) Y_i, \quad i = 1,2,\cdots,k,$$

$$\underset{n\times p}{Y} = (Y_1' \, Y_2' \cdots Y_k')', \quad n = \sum_{i=1}^{k} n_i,$$

$$\bar{y} = \frac{1}{n} Y' \mathbb{1}, \quad S = Y'\left(I - \frac{1}{n} \mathbb{1}\mathbb{1}'\right) Y.$$

于是检验(2.3)的统计量是

$$\Lambda = \frac{|S_1 + \cdots + S_k|}{|S|}, \tag{2.4}$$

Λ 的分布是$\Lambda(p, n-k, k)$，这一分布的表可以在文献[1]的附录

中找到. 当 $k=2$ 时,就是两总体的比较,此时(2.4)相应的 Λ 的分布可以从下式

$$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{n-p}{p}\sim F(2p,2(n-p))$$

得到,或直接使用统计量 $\dfrac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}$.

对于成分数据向量 x,它的特殊一些的均值检验问题是有关部分分量和子成分的检验,现在我们来论述这类问题的背景和它的意义.

首先,关于部分分量的检验和子成分的检验是不同的,但又是有联系的. 关于部分分量的检验,例如一种橡胶,它是由天然橡胶以及其他各种人工合成的橡胶配方混炼而成的,人们关注的是天然橡胶含量的百分比是否是规格中规定的量,这时要检验的是成分中某些分量是否为特定的值. $\underset{(p+1)\times 1}{x}$ 表示成分向量,它的期望值 Ex 记为 θ,将 θ 分为两段 $\theta=\begin{pmatrix}\theta_{(1)}\\ \theta_{(2)}\end{pmatrix}$,要检验的统计问题是

$$H_0:\theta_{(1)}=\theta_{(1)0} \tag{2.5}$$

是否成立,$\theta_{(1)0}$是一给定的向量. 这一类问题是检验部分分量问题. 某一地区的农产量分粮食、经济、油料等各种类型,随着经济的发展,粮食作物中大米、小麦、薯类等占的比例是否有变化,这时整个农作物的产值会变化,经济作物与粮食、油料等作物的比例会变化,粮食作物的总产量也会变化,我们关心的是粮食作物中大米、小麦、薯类等比例是否有变化. 如果用 x 表示农作物中各种作物的播种面积所占的比例,$\theta=Ex$,于是 θ 中的一部分设为 $\theta_{(1)}$就是粮食作物播种面积的比例,但我们并不关注 $\theta_{(1)}$这个部分是否有变化,而是关心 $\theta_{(1)}$中各成分的相对比例——子成分

$$S_{(1)}=\theta_{(1)}(\mathbb{1}'\theta_{(1)})^{-1} \tag{2.6}$$

是否有变化,因此要检验的问题是

$$H_0:S_{(1)}=S_{(1)0}, \tag{2.7}$$

其中 $S_{(1)0}$是一指定的常数向量.很明显,(2.7)相应的问题,仍然是一个成分向量的均值检验,注意到加法逻辑正态分布的性质,它的子成分仍然是加法逻辑正态分布,因此这一类的检验问题仍然归结为例 2.1,例 2.2,例 2.3 的情形,自然就可以解决.但是(2.5)的问题是不同的,它不能化成一个子成分的问题,因为 $\theta_{(1)}$ 与 $\theta_{(1)}^*$ 可以不同,但 $S_{(1)}=\theta_{(1)}(\mathbb{1}'\theta_{(1)})^{-1}$与 $S_{(1)}^*=\theta_{(1)}^*(\mathbb{1}'\theta_{(1)}^*)^{-1}$可以是相同的.因此(2.5)成立时,一定有(2.7);然而(2.7)成立时,不一定有(2.5).

如果 $\theta_{(2)}$中有一个成分不变,而且这一情况是已知的,则有办法把它作为 x_0 来处理,这时就相当于对 y 的总体$N(\mu,\Sigma)$,只检验 μ 的一部分,即

$$\mu = \begin{pmatrix}\mu_{(1)}\\ \mu_{(2)}\end{pmatrix},$$

只问

$$H_0: \mu_{(1)} = \mu_{(1)0}$$

是否成立,$\mu_{(1)0}$是一固定的向量.这一类问题在一般多元分析的书中已有讨论,这里就不进行这方面的论述.下面将这个问题转换为乘法逻辑正态分布的检验问题.

在第二章中,我们讨论了乘法逻辑正态分布,即对成分向量 $x=(x_0,x_1,\cdots,x_p)'$假定

$$y_i = \ln\left(x_i\Big/\left(1-\sum_{j=0}^{i}x_j\right)\right) \quad i=1,2,\cdots,p$$

是正态分布.因此要检验某一段分量的期望值是否为指定的向量时,总可以认为它是$(x_0,\cdots,x_k)$组成的,因此从乘法逻辑正态分布来看,就相当于检验$(y_1,\cdots,y_k)$是否来自期望值向量为指定向量的总体,这就还原为一个正态总体的检验问题.然而困难的是,对 $\theta=Ex$ 的一个命题并不能直接表示为对Ey 的一个命题,这一点我们从上节的参数估计中已清楚地看到了.

对乘法逻辑正态分布,已知 $y\sim N(\mu,\Sigma)$,注意此时 y 与 x 的关系式是

$$\begin{cases} x_i = e^{y_i} \Big/ \left[\prod_{j=1}^{i} (1 + e^{y_j}) \right], & i = 1,2,\cdots,p, \\ x_0 = \left[\prod_{j=1}^{p} (1 + e^{y_j}) \right]^{-1} \end{cases}$$

或

$$y_i = \ln x_i - \ln\left(1 - \sum_{j=0}^{i} x_i\right), \quad i = 1,2,\cdots,p .$$

因此,从

$$\begin{cases} Ex_i = E\left(e^{y_i} \left[\prod_{j=1}^{i} (1 + e^{y_j})^{-1} \right]\right), & i = 1,2,\cdots,p, \\ Ex_0 = E \prod_{j=1}^{p} (1 + e^{y_j})^{-1} \end{cases}$$

或

$$Ey_i = E\ln x_i - E\ln\left(1 - \sum_{j=0}^{i} x_i\right), \quad i = 1,2,\cdots,p .$$

都只能得到 $\theta = Ex$ 与 $\mu = Ey$ 之间的近似关系式,并不能直接给出一个完全等价的命题.

从这个讨论可以看出,加法逻辑正态分布,乘法逻辑正态分布对于处理变换后的 y 而言,y 的期望、协差阵都已和正态相同,但从要讨论的成分向量 x 的期望值 $\theta = Ex$ 与协方差矩阵 $\mathscr{K}$ 而言,变换后的 y 并不能给出明确而有效的处理方法,这是成分向量统计分析的困难之处.

§3. 主分量分析、典型相关分析

注意到加法逻辑正态分布的特性是处理成分向量经变换的正态变量 y,因此对于 y 可以进行主分量分析.这与通常正态分布中的主分量分析没有什么不同,这里我们利用加法逻辑正态的特性,引入成分向量 x 的另一种变换,使它在处理主分量分析时带来一些方便.

设 $x=\begin{pmatrix}x_0\\ \vdots\\ x_p\end{pmatrix}$是成分向量，

$$y=(-\mathbb{1}\quad I_p)\ln x=\begin{pmatrix}\ln(x_1/x_0)\\ \vdots\\ \ln(x_p/x_0)\end{pmatrix}\sim N\left(\mu,\sum\right).$$

令

$$z=\ln x-\mathbb{1}\ln g(x)=\begin{pmatrix}\ln(x_0/g(x))\\ \vdots\\ \ln(x_p/g(x))\end{pmatrix},$$

其中

$$g(x)=\left(\prod_{i=0}^{p}x_i\right)^{\frac{1}{p+1}},$$

因此

$$\underset{(p+1)\times 1}{z}=\left(I-\frac{1}{p+1}\mathbb{1}\,\mathbb{1}'\right)\ln x\triangleq Q\ln x\ . \tag{3.1}$$

注意到$\left(I-\frac{1}{p+1}\mathbb{1}\,\mathbb{1}'\right)\mathbb{1}=0$，因此 $Q\,\mathbb{1}\ln x_0=0$，于是

$$z=Q\ln x-Q\,\mathbb{1}\ln x_0=Q\ln(x/x_0)=Q\begin{pmatrix}0\\ y\end{pmatrix},$$

因此 z 也是正态分布，但是它是退化的正态分布，z 的协方差矩阵 $\sum_z$ 的行列式是 0. 从上述表达式可知 $z=\theta_* y$，因此 $z\sim N(\theta_*\mu,\theta_*\sum\theta'_*)$，其中

$$\theta_*=\begin{pmatrix}-\frac{1}{p+1}\mathbb{1}'_p\\ I_p-\frac{1}{p+1}\mathbb{1}\,\mathbb{1}'\end{pmatrix}.$$

由于 z 本身的定义对于成分向量 x 中各个坐标分量具有对称性，因此对它进行主分量分析就更能反映成分的特性.

容易证明下面 y 与 z 的相互表示的公式，这就留作练习. 记

$$\underset{p\times(p+1)}{F} = (-\mathbb{1} \quad I_p) , \tag{3.2}$$

于是

$$\underset{(p+1)\times p}{F^+} = F'(FF')^+ = F'(FF')^{-1} = \begin{pmatrix} -\mathbb{1}' \\ I_p \end{pmatrix}\left(I_p - \frac{1}{p+1}\mathbb{1}\mathbb{1}'\right)$$

$$= \begin{bmatrix} -\dfrac{1}{p+1}\mathbb{1}' \\ I_p - \dfrac{1}{p+1}\mathbb{1}\mathbb{1}' \end{bmatrix} = \begin{pmatrix} 0 \\ I_p \end{pmatrix} - \frac{1}{p+1}\mathbb{1}_{p+1}\ \mathbb{1}_p' = Q_* . \tag{3.3}$$

并且有

$$z = F^+ y, \quad y = Fz . \tag{3.4}$$

现在来讨论 z 的主分量,也就是对 z 的协方差矩阵

$$\sum\nolimits_z = \mathrm{Var}(z) = \mathrm{Var}(F^+ y) = F^+ \sum F^{+'}$$

进行谱分解.

设 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{p+1}$ 是 $\sum_z$ 的 $p+1$ 个特征根,由于 F^+ 的秩是 p,因此可以断定 $\lambda_{p+1}=0$,实际上只有 p 个主分量.记非零特征根 λ_i 相应的特征向量为 t_i,于是有下列命题.

命题 3.1 设 t_i 是 $\lambda_i \neq 0$ 相应的特征向量,则

$$\mathbb{1}' t_i = 0 . \tag{3.5}$$

证明 今 $\lambda_i t_i = \sum_z t_i = Q\mathrm{Var}(\ln x)Q t_i$,由于 $Q\,\mathbb{1}=0$,因此

$$\lambda_i \mathbb{1}' t_i = \mathbb{1}' Q \mathrm{Var}(\ln x) Q t_i = 0 ,$$

今 $\lambda_i \neq 0$,就得 $\mathbb{1}' t_i = 0$.

命题 3.2 设 t_i 为 $\lambda_i \neq 0$ 相应的特征向量,则它相应的主分量 c_i 有表达式

$$c_i = t_i' \ln x \tag{3.6}$$

证明 这是因为

$$c_i = t_i' z = t_i' Q \ln x$$

$$= t_i' \left(I - \frac{1}{p+1}\mathbb{1}\mathbb{1}'\right) \ln x$$

$$= t_i' \ln x .$$

从(3.5)及(3.6)就知道,$\ln x$ 的对比 $t_i' \ln x$ 就是 z 的主分量,也就是成分向量 x 的对数对比.这些主分量也都是正态分布,它们也都是 y 的线性函数.

从上面的讨论,利用成分向量的样本矩阵 $\underset{n\times(p+1)}{X} = \begin{pmatrix} x'_{(1)} \\ \vdots \\ x'_{(n)} \end{pmatrix}$,

加工成 $\ln X = (\ln x_{ij}) = \begin{pmatrix} \ln x'_{(1)} \\ \vdots \\ \ln x'_{(n)} \end{pmatrix}$,再形成

$$\underset{n\times(p+1)}{Z} = (\ln X)Q ,$$

然后对

$$S_z = \frac{1}{n-1} Z' \left(I_n - \frac{1}{n} \mathbb{1}\mathbb{1}' \right) Z$$

求特征根、特征向量,以特征向量为系数,就求得各个主分量.

对于成分向量 x,当 x 遵从加法逻辑正态分布时,也可以考虑它的对数 $\ln x$ 分两组或多组时的典型相关分析,如直接对 y 讨论,这时 x_0 的地位有些特殊,缺乏各分量之间的对称性,所以我们还是对 z 进行讨论,z 对于各个分量是完全对称的.

将 z 分成两组,

$$z = \begin{pmatrix} z_{(1)} \\ z_{(2)} \end{pmatrix} = Q \ln x = \begin{pmatrix} Q_1 \ln x \\ Q_2 \ln x \end{pmatrix} \begin{matrix} k+1 \\ \cdot \\ p-k \end{matrix},$$

其中

$$\underset{(k+1)\times(p+1)}{Q_1} = \left(I_{k+1} - \frac{1}{p+1} \mathbb{1}\mathbb{1}' \quad - \frac{1}{p+1} \mathbb{1}\mathbb{1}' \right),$$

$$\underset{(p-k)\times(p+1)}{Q_2} = \left(- \frac{1}{p+1} \mathbb{1}\mathbb{1}' \quad I_{p-k} - \frac{1}{p+1} \mathbb{1}\mathbb{1}' \right).$$

现在利用判别信息量来导出两个随机向量间的相关性度量.我们先导出非退化正态随机向量的相关性度量,然后将它用于退化正态分布.相关性度量用联合密度 $p(x,y)$ 与独立时边缘密度

的乘积 $f(x)g(y)$的判别信息量来反映,其中

$$f(x) = \int p(x,y)\mathrm{d}y, \quad g(y) = \int p(x,y)\mathrm{d}x,$$

判别信息量是

$$I(x,y) = \iint p(x,y)\ln \frac{p(x,y)}{f(x)g(y)}\mathrm{d}x\mathrm{d}y \ . \qquad (3.7)$$

如果

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \begin{bmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{bmatrix}\right],$$

则自然有

$$x \sim N\left(\mu_x, \sum\nolimits_{xx}\right), y \sim N\left(\mu_y, \sum\nolimits_{yy}\right) \ .$$

用(3.7)式就可以算出

$$\begin{aligned}
I(x,y) = E\Bigg[& -\frac{1}{2}(x' - Ex'y' - Ey')\begin{bmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{bmatrix}^{-1}\begin{pmatrix} x - Ex \\ y - Ey \end{pmatrix} \\
& + \frac{1}{2}(x - Ex)' \sum\nolimits_{xx}^{-1}(x - Ex) \\
& + \frac{1}{2}(y - Ey)' \sum\nolimits_{yy}(y - Ey)\Bigg] \\
& - \left(\frac{1}{2}\ln\begin{vmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{vmatrix} - \frac{1}{2}\ln\left|\sum\nolimits_{xx}\right| - \frac{1}{2}\ln\left|\sum\nolimits_{yy}\right|\right) \\
= -\frac{1}{2}\ln & \left(\begin{vmatrix} \sum_{xx} & \sum_{xy} \\ \sum_{yx} & \sum_{yy} \end{vmatrix} \Big/ \left(\left|\sum\nolimits_{xx}\right| \left|\sum\nolimits_{yy}\right|\right)\right) \\
= -\frac{1}{2}\ln & \left| I - \sum\nolimits_{yy}^{-1}\sum\nolimits_{yx}\sum\nolimits_{xx}^{-1}\sum\nolimits_{xy}\right| \ .
\end{aligned}$$

因此相关性度量就与矩阵

$$\sum\nolimits_{yy}^{-\frac{1}{2}}\sum\nolimits_{yx}\sum\nolimits_{xx}^{-1}\sum\nolimits_{xy}\sum\nolimits_{yy}^{-\frac{1}{2}} \qquad (3.8)$$

的特征根 λ_i 有关,这些特征根的 $\lambda_i^{\frac{1}{2}}$ 就是典型相关系数.对于退化的正态分布,只要将(3.8)中的逆矩阵改成"+"号逆,即考虑矩阵

$$\sum_{yy}^{+\frac{1}{2}}\sum_{yx}\sum_{xx}^{+}\sum_{xy}\sum_{yy}^{+\frac{1}{2}} \tag{3.9}$$

就可以了.这样我们就导出了两组成分分量之间的典型相关系数,它就是矩阵

$$(Q_2VQ_2')^{+\frac{1}{2}}(Q_2VQ_1')(Q_1VQ_1')^{+}(Q_1VQ_2')(Q_2VQ_2')^{+\frac{1}{2}}$$

的非零特征根,其中

$$V = \mathrm{Var}(\ln x).$$

从成分向量 x 的样本矩阵 $\underset{n\times(p+1)}{X}$ 出发,加工成 $\underset{n\times(p+1)}{Z} = (\ln X)Q$,将 Z 相应分块,分成

$$\underset{n\times(p+1)}{Z} = (\underset{n\times(k+1)}{Z_1} \quad \underset{n\times(p-k)}{Z_2}),$$

再加工成

$$L_{11} = Z_1'\left(I_n - \frac{1}{n}\mathbb{1}\mathbb{1}'\right)Z_1,$$

$$L_{12} = Z_1'\left(I_n - \frac{1}{n}\mathbb{1}\mathbb{1}'\right)Z_2 = L_{21}',$$

$$L_{22} = Z_2'\left(I_n - \frac{1}{n}\mathbb{1}\mathbb{1}'\right)Z_2,$$

求

$$L_{22}^{+}L_{21}L_{11}^{+}L_{12} \text{ 或 } L_{22}^{+\frac{1}{2}}L_{21}L_{11}^{+}L_{12}L_{22}^{+\frac{1}{2}}$$

的非零特征根,就能得典型相关系数,典型变量的求法就与正态时非退化的情形相仿,这里就不再叙述了.

§4. 子成分的独立性

从成分向量 x 的分布是加法逻辑正态分布,就知道它的任一子成分向量的分布也是加法逻辑正态.因为子成分的任一对数对比一定是 x 的一个对数对比.

现在来讨论子成分的独立性检验.将成分向量分成两段,$x=\begin{pmatrix}x_{(1)}\\x_{(2)}\end{pmatrix}\begin{matrix}p_1\\p_2\end{matrix}$,$p_1+p_2=p+1$,先讨论两个子成分

$$S_{(i)}=x_{(i)}(\mathbb{1}'x_{(i)})^{-1},i=1,2, \tag{4.1}$$

的独立性.

引理 4.1 当 $p_i\geqslant 2$ 时,x 服从加法逻辑正态分布,则有 $S_{(1)}$ 与 $S_{(2)}$ 相互独立的充分必要条件是

$$P_1V_{12}P_2=0,V=\begin{pmatrix}V_{11}&V_{12}\\V_{21}&V_{22}\end{pmatrix}, \tag{4.2}$$

其中

$$P_i=I_{p_i}-\frac{1}{p_i}\mathbb{1}\mathbb{1}',i=1,2,V=\mathrm{Var}(\ln x).$$

证明 $S_{(1)}$ 与 $S_{(2)}$ 独立的充分必要条件是向量 $\ln S_{(1)}$ 与 $\ln S_{(2)}$ 相互独立,这也等价于 $P_1\ln S_{(1)}$ 与 $P_2\ln S_{(2)}$ 相互独立.由于 $P_i\ln S_{(i)}$ 的正态性,独立的充分必要条件是不相关.今

$$0=\mathrm{Cov}(P_1\ln S_{(1)},P_2\ln S_{(2)})=P_1\mathrm{Cov}(\ln S_{(1)},\ln S_{(2)})P_2,$$

注意到 $\ln S_{(i)}=\ln x_{(i)}-\mathbb{1}\ln(\mathbb{1}'x_{(i)})$,因此有

$$P_i\ln S_{(i)}=P_i(\ln x_{(i)}-\mathbb{1}\ln(\mathbb{1}'x_{(i)}))=P_i\ln x_{(i)},$$

于是就得到(4.2).

现在要将(4.2)式用 $y=(-\mathbb{1}\ \ I_p)\ln x$ 的协方差矩阵Σ表示出来,因此要将Σ分块,注意到 x_0 在 y 中的特殊位置,这样必须分几种情形逐一的讨论.

将 x 分成k 段,$x=(x'_{(1)},x'_{(2)},\cdots,x'_{(k)})'$,$x_{(i)}$是 p_i 维向量,$p_i\geqslant 2$,且 $\sum_{i=1}^k p_i=p+1$,相应的

$$\mathrm{Var}(\ln x)=V=\begin{pmatrix}V_{11}&V_{12}&\cdots&V_{1k}\\\vdots&\vdots&\ddots&\vdots\\V_{k1}&V_{k2}&\cdots&V_{kk}\end{pmatrix}\begin{matrix}p_1\\\vdots\\p_k\end{matrix}, \tag{4.3}$$

记 $P_i=I_{p_i}-\frac{1}{p_i}\mathbb{1}\mathbb{1}',i=1,\cdots,k.$

当 $i\neq 1$ 时，$P_i \ln S_{(i)} = P_i \ln x_{(i)} = P_i(\ln x_{(i)} - \mathbb{1}\ln x_0)$
$$= P_i y_{(i)}.$$
注意，$y=(-\mathbb{1}\ I_p)\ln x$ 也与 x 相应的 k 段，但 $y_{(1)}$ 是 p_1-1 维，$y=(y'_{(1)}, y'_{(2)}, \cdots, y'_{(k)})'$，$y$ 相应的协方差矩阵也相应分成 $k\times k$ 的分块形式，即

$$\mathrm{Var}(y) = \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1k} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{k1} & \Sigma_{k2} & \cdots & \Sigma_{kk} \end{pmatrix} \begin{matrix} p_1-1 \\ p_2 \\ \vdots \\ p_k \end{matrix} \quad (4.4)$$

当 $i=1$ 时，

$$P_1 \ln S_{(1)} = P_1 \ln x_{(1)} = P_1(\ln x_{(1)} - \mathbb{1}\ln x_0)$$
$$= P_1 \begin{pmatrix} 0 \\ y_{(1)} \end{pmatrix} = \begin{pmatrix} -p_1^{-1}\mathbb{1}' \\ I_{p_1-1} - p_1^{-1}\mathbb{1}\mathbb{1}' \end{pmatrix} y_{(1)}.$$

有了这些准备，就可以用引理 4.1 来给出由 Σ 表示的充分必要条件.

定理 4.1 设成分向量 x 服从加法逻辑正态分布，相应的 $y=(-\mathbb{1}\ I_p)\ln x \sim N(\mu, \Sigma)$，则将 x 分成 k 段 $x_{(1)}, \cdots, x_{(k)}$ 后，就有如下的结论：

$$\begin{cases} \text{当 } i\neq 1, j\neq 1 \text{ 时}, S_{(i)}, S_{(j)} \text{ 独立} \Leftrightarrow P_i \Sigma_{ij} P_j = 0 \\ \text{当 } i\neq 1, j=1 \text{ 时}, S_{(i)}, S_{(1)} \text{ 独立} \Leftrightarrow P_i \Sigma_{i1} = 0 \end{cases} \quad (4.5)$$

$i,j=2,3,\cdots,k, i\neq j$.

这一证明是简单的，因为当 $i\neq 1, j\neq 1$ 时

$$0 = \mathrm{Cov}(P_i \ln S_{(i)}, P_j \ln S_{(j)})$$
$$= \mathrm{Cov}(P_i y_{(i)}, P_j y_{(j)}) = P_i \Sigma_{ij} P_j$$

当 $i\neq 1$ 时，

$$0 = \mathrm{Cov}(P_i \ln S_{(i)}, P_1 \ln S_{(1)})$$

$$= \mathrm{Cov}\left[P_i y_{(i)}, P_1 \begin{bmatrix} 0 \\ y_{(1)} \end{bmatrix}\right]$$

$$= P_i \Sigma_{i1}\left(-\frac{1}{p_1}\mathbb{1}, I_{p_1-1} - \frac{1}{p_1}\mathbb{1}\mathbb{1}'\right)$$

即

$$0 = \left(P_i \Sigma_{i1} \mathbb{1}\left(-\frac{1}{p_1}\right), P_i \Sigma_{i1} - P_i \Sigma_{i1} \mathbb{1}\mathbb{1}' p_1^{-1}\right)$$

上式成立的充分必要条件是 $P_i \Sigma_{i1} = 0$.

从定理 4.1 知道,要检验子成分的独立性,就引出正态总体协方差阵结构的假设检验问题. 现在我们从(4.5)导出相应的统计量.

$P_i \Sigma_{ij} P_j = 0$,表示 $P_i y_{(i)}$ 与 $P_j y_{(j)}$ 不相关,由于 $y_{(i)}, y_{(j)}$ 是正态变量,也就是 $P_i y_{(i)}$ 与 $P_j y_{(j)}$ 相互独立. 这也就是 $P_i y_{(i)}$ 与 $P_j y_{(j)}$ 相应的典型相关系数全为 0. 注意到样本相应的典型相关系数可以 y 的样本协方差阵的分块表示后的矩阵来导出,这样就可以获得有关的统计量.

将 y 的样本矩阵 $Y_{n \times p}$ 也分块写出,于是

$$\hat{\Sigma} = \frac{1}{n-p} Y' \left(I_n - \frac{1}{n}\mathbb{1}\mathbb{1}'\right) Y$$

$$= \frac{1}{n-p}(L_{ij}),$$

其中

$$L_{ij} = Y_i' \left(I_n - \frac{1}{n}\mathbb{1}\mathbb{1}'\right) Y_j,$$

$$Y = \begin{pmatrix} Y_1 & Y_2 & \cdots & Y_k \end{pmatrix}$$
$$\quad\; p_1 - 1 \quad p_2 \quad \cdots \quad p_k.$$

样本的典型相关系数的平方是矩阵 $\hat{M}_{ij}$ 的特征根,$\hat{M}_{ij} = L_{ii}^{-1} L_{ij} \cdot L_{jj}^{-1} L_{ji}$. 由于 $P_i y_{(i)}$ 的样本矩阵是 $Y_i P_i$,因此 $P_i y_{(i)}$ 与 $P_j y_{(j)}$ 的典型相关系数的平方是矩阵

$$(P_i L_{ii} P_i)^+ P_i L_{ij} P_j (P_j L_{jj} P_j)^+ P_j L_{ji} P_i$$

的特征根. 当 $P_iy_{(i)}$ 与 $P_jy_{(j)}$ 相互独立时, 所有典型相关系数均为 0, 因此它们的平方也是 0, 它的平方和也是 0, 这就导出

$$t = \mathrm{tr}(P_iL_{ii}P_i)^+ (P_iL_{ij}P_j)(P_jL_{jj}P_j)^+ P_jL_{ji}P_i$$

是一个合适的检验统计量.

注意到 $P_i\Sigma_{i1}=0$ 也就是 $P_iy_{(i)}$ 与 $y_{(1)}$ 独立, 因此相应的统计量是

$$t = \mathrm{tr}(P_iL_{ii}P_i)^+ (P_iL_{i1})L_{11}^+L_{1i}P_i.$$

它们精确分布并不好求, 但它们的渐近分布是可以得到的, 这样检验的问题也就解决了.

§5. 回归分析

成分数据的回归分析有两类: 一类是以成分向量为因变量, 考虑其余的变量为自变量; 一类是以成分向量为自变量, 以其他变量为因变量, 现在分别讨论如下.

以成分向量为因变量的回归问题, 就是考虑别的变量对成分的影响, 例如家庭的收入对家庭消费的各种成分会有影响, 家庭消费可以分为食品、衣着、交际、文化教育、其他等项, 很明显, 收入不同时, 上述成分占的比例是不同的, 因此这一类回归问题是有意义的, 一般说来, 它是多对多的回归.

直接将成分作为因变量来建立回归方程, 由于成分 x_i 的值在 (0,1) 区间内, 对回归函数的选择带来很多困难, 因此选用 $y_i=\ln\dfrac{x_i}{x_0}, i=1,2,\cdots,p$ 作为因变量有很多方便, 一是 y_i 的值可以在 $(-\infty,\infty)$ 内变化, 对函数的选择带来方便; 二是 y_i 可以认为是服从正态分布的, 只要成分向量 x 是遵从加法逻辑正态就可以了. 这样用 y_i 作因变量就与普通的多元回归没有什么区别.

值得注意的是一类由基向量引导的回归问题. 例如, 一个家庭有各种各样的收入, 也有各种各样的支出, 因此收入与支出是两个随机向量, 考虑到它们的分量都不是负的, 因此用联合对数正态来

描述还是合适的.我们用 u 向量描述各种收入,用 ω 向量描述各种支出,它们的联合分布为

$$\begin{matrix} q \\ p+1 \end{matrix}\begin{pmatrix} \ln u \\ \ln \omega \end{pmatrix} \sim N\left[\begin{bmatrix} \theta_u \\ \theta_\omega \end{bmatrix}, \begin{bmatrix} \Omega_{uu} & \Omega_{u\omega} \\ \Omega_{\omega u} & \Omega_{\omega\omega} \end{bmatrix}\right].$$

由 ω 导出支出的成分向量 $x=\omega(\mathbb{1}'\omega)^{-1}$,相应的 $y_i=\ln(x_i/x_0),i=1,2,\cdots,p$ 是正态分布.

此时从 $\ln u$,$\ln\omega$ 可以导出 $\ln u$ 与 y 的联合分布,然后利用条件期望和条件方差,求出相应回归方程的函数形式与加权最小二乘的权矩阵,下面我们就来推导这些公式.

注意到 $\omega=\begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_p \end{bmatrix}$,且 $x=\omega(\mathbb{1}'\omega)^{-1}$,

$$\begin{aligned} y_i &= \ln(x_i/x_0) = \ln(\omega_i(\mathbb{1}'\omega)^{-1}/\omega_0(\mathbb{1}'\omega)^{-1}) \\ &= \ln\omega_i - \ln\omega_0, \quad i=1,2,\cdots,p. \end{aligned}$$

因此

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = (-\mathbb{1} \quad I_p)\ln\omega \triangleq F\ln\omega,$$

于是

$$\begin{aligned} Ey &= F\theta_\omega, \mathrm{Var}(y) = F\Omega_{\omega\omega}F', \\ \mathrm{Cov}(y,\ln u) &= \mathrm{Cov}(F\ln\omega,\ln u) \\ &= F\Omega_{\omega u}. \end{aligned}$$

这样我们就得到 y 对 u 的条件期望和条件方差的公式(直接引用正态分布的结论)是:

$$\begin{aligned} E\{y \mid u\} &= E\{y \mid \ln u\} \\ &= Ey + \mathrm{Cov}(y,\ln u)\mathrm{Var}(\ln u)(\ln u - \theta_u) \\ &= F\theta_\omega + F\Omega_{\omega u}\Omega_{uu}^{-1}(\ln u - \theta_u) \\ &= FE\{\ln\omega \mid \ln u\}, \end{aligned}$$

$$\begin{aligned}\mathrm{Var}\{y \mid u\} &= \mathrm{Var}\{y \mid \ln u\} = \mathrm{Var}\{F\ln\omega \mid \ln u\} \\ &= F\mathrm{Var}\{\ln\omega \mid \ln u\}F' \\ &= F(\Omega_{\omega\omega} - \Omega_{\omega u}\Omega_{uu}^{-1}\Omega_{u\omega})F'.\end{aligned}$$

所以成分向量 x 对 u 的回归方程，用 y 来表示时，它是 u 的对数的线性函数，它的权矩阵也可以通过上述公式求得.

现在来讨论第二类问题，它的表达形式自然会与混料设计的内容有联系，在这里，我们先不去讨论设计问题，只讨论回归问题及相关性的度量. 用成分向量 x 作为自变量，与用它的变换 $y_i = \ln x_i / x_0, i = 1,2,\cdots,p$ 作为自变量相比，后者有很多方便，我们用 u 表示问题中要考虑的因变量，因此可以从这样的前提出发，就是 u 与 y 是联合正态分布的，这样 u 对 y 的回归函数的形式可以从条件期望的表达式中得到. 因此，同上一个类型的问题相仿，u 是 y 的线性函数，也就是说，u 是成分向量 x 对数的线性函数.

从这个角度来看混料试验的设计与分析，对于优良设计的选择，可以有另一种考虑方法，下面用一个例子来说明.

例 5.1 混料试验问题的简单情形.

在橡胶工艺中，要设计生产的橡胶在性能上达到某种要求，就需要考虑配方，配方是天然胶用多少(用百分比描述)，人工胶用多少，填充剂用多少，还有工艺水平的选择等等. 原材料各种成分的比例，就是很自然的一个成分向量，我们用 u 表示某个性能，x_0, $x_1, \cdots, x_p$ 表示成分，于是有关系式：

$$\begin{cases} u_\alpha = a + \sum_{i=0}^{p} b_i \ln x_{\alpha i} + \varepsilon_\alpha, \alpha = 1,2,\cdots,n, \\ \text{且 } 0 < x_{\alpha i} < 1, i = 0,1,2,\cdots,p, \\ \sum_{i=0}^{p} x_{\alpha i} = 1, \alpha = 1,\cdots,n, \end{cases} \tag{5.1}$$

其中

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, E\varepsilon = 0, \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

(5.1)式相应的设计矩阵

$$\underset{n\times(p+2)}{H}=\begin{pmatrix}1 & \ln x_{10} & \cdots & \ln x_{1p}\\ \vdots & \vdots & & \vdots\\ 1 & \ln x_{n0} & \cdots & \ln x_{np}\end{pmatrix}.$$

因此

$$H'H=\begin{pmatrix}\mathbb{1}'\\ \ln x'_{(0)}\\ \vdots\\ \ln x'_{(p)}\end{pmatrix}(\mathbb{1}\ \ln x_{(0)}\cdots\ln x_{(p)})$$

$$=\begin{pmatrix} n & \sum_{\alpha=1}^{n}\ln x_{\alpha 0} & \cdots & \sum_{\alpha=1}^{n}\ln x_{\alpha p}\\ \sum_{\alpha=1}^{n}\ln x_{\alpha 0} & & & \\ \vdots & & \sum_{\alpha=1}^{n}(\ln x_{\alpha i})(\ln x_{\alpha j}) & \\ \sum_{\alpha=1}^{n}\ln x_{\alpha p} & & & \end{pmatrix}.$$

注意到 $H'H$ 的表达式中只与试验次数 n 与成分向量 n 个点的选择有关.从试验设计的优良性来看,有 D 最优、A 最优与 λ 最优这些准则,而这些准则的优化目标都是 $H'H$ 的函数.

D 最优:要求 $H'H$ 的行列式 $|H'H|$ 最大;

A 最优:要求 $\operatorname{tr}(H'H)^{-1}$ 达到最小;

λ 最优:要求 $H'H$ 的最小特征根达到最大.注意到 $\operatorname{tr}(H'H)^{-1}$ 是 $H'H$ 的逆矩阵的特征根之和,用 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_{p+2}$ 表示 $H'H$ 的特征根,则有

$$\operatorname{tr}(H'H)^{-1}=\sum_{i=1}^{p+2}\lambda_i^{-1}\geqslant(p+2)^2\Big(\sum_{i=1}^{p+2}\lambda_i\Big)^{-1}$$
$$=(p+2)^2(\operatorname{tr}H'H)^{-1}.$$

当 n 给定后 $\operatorname{tr}H'H$ 尽可能大,也就是使 $\operatorname{tr}(H'H)^{-1}$ 的下确界尽可

能小，这样，优良设计的选择都与 $H'H$ 的特征根有关，而且都要求 $H'H$ 在某种意义下尽可能地“大”. 注意到 $H'H$ 是向量的内积矩阵，$|H'H|$ 反映的是由这 $p+2$ 个 n 维向量形成的多面体体积，$\mathrm{tr}H'H$是这 $p+2$ 个向量的各个长度平方和，$H'H$ 的最小特征根反映的是“椭球”的短轴，所以优良设计的几何意义是明显的，要求 $p+2$个向量尽可能“分散”.

现在我们看一个特殊的优良设计的解，这个求解的方法对我们具体处理设计问题是有启发的. 设已做了 n 次试验，相应的设计矩阵 H 已知，现在要添加一个试验，问这个试验点的成分向量该如何选择？我们用 H_* 表示添加一次试验后形成的设计矩阵，于是 H 与 H_* 的关系可以表示成：

$$H_* = \begin{pmatrix} & H & \\ 1 & \ln x_0^* \cdots \ln x_p^* \end{pmatrix} \triangleq \begin{pmatrix} & H \\ 1 & (\ln x^*)' \end{pmatrix},$$

其中 $x_0^*, \cdots, x_p^*$ 就是待选的成分向量 x^*.

如果考虑 D 最优，则选 x^* 使 $|H'_* H_*|$ 达到最大. 此时

$$H'_* H_* = \begin{pmatrix} H' & 1 \\ & \ln x^* \end{pmatrix} \begin{pmatrix} & H \\ 1 & (\ln x^*)' \end{pmatrix}$$

$$= H'H + \begin{pmatrix} 1 \\ \ln x^* \end{pmatrix} (1 \ (\ln x^*)'),$$

于是

$$|H'_* H_*| = \left| H'H + \begin{pmatrix} 1 \\ \ln x^* \end{pmatrix} (1 \ (\ln x^*)') \right|$$

$$= \left[1 + (1 \ (\ln x^*)')(H'H)^{-1} \begin{pmatrix} 1 \\ \ln x^* \end{pmatrix} \right] |H'H|,$$

所以只要选 x^* 使

$$f = (1 \ \ (\ln x^*)')(H'H)^{-1} \begin{pmatrix} 1 \\ \ln x^* \end{pmatrix} \qquad (5.2)$$

达到最大，x^* 就是一个优良的设计. 这是一个有约束条件的极值问题，为了能比较清楚看出对 x^* 的依赖，记 $(H'H)^{-1} = \begin{pmatrix} a & b' \\ b & C \end{pmatrix}$,

$C=(c_{ij})$,于是

$$f = a + 2b'\ln x^* + (\ln x^*)'C\ln x^*$$
$$= a + 2\sum_{i=0}^{p} b_i \ln x_i^* + \sum_{i,j=0}^{p} c_{ij}(\ln x_i^*)(\ln x_j^*),$$

约束条件是 $x_i^* \geqslant 0, i=0,1,\cdots,p, \sum_{i=0}^{p} x_i^* = 1$. 用拉氏乘子法,得

$$\frac{2(f-(\lambda\sum x_i^*-\lambda))}{2x_i^*}$$
$$=2b_i\frac{1}{x_i^*}+2\sum_{j=0}^{p}c_{ij}\frac{1}{x_i^*}\ln x_j^*-\lambda, i=0,1,\cdots,p.$$

让上式为0,用矩阵的形式写出,记 $(x^*)^{-1}=\begin{pmatrix}\frac{1}{x_0^*}\\ \vdots\\ \frac{1}{x_p^*}\end{pmatrix}$,则有方程

$$\begin{cases}\begin{pmatrix}\frac{1}{x_0^*} & & 0\\ & \ddots & \\ 0 & & \frac{1}{x_p^*}\end{pmatrix}C\ln x^* = \frac{\lambda}{2}\mathbb{1}-\begin{pmatrix}b_0/x_0^*\\ \vdots\\ b_p/x_p^*\end{pmatrix}\\ x_i^*\geqslant 0, i=0,1,2,\cdots,p, \sum_{i=0}^{p}x_i^*=1.\end{cases} \tag{5.3}$$

很明显,(5.3)是一个非线性方程组,它的解无法用显式表示,但可以用迭代法求解.

在(5.3)式中,约束条件给求解带来了一些麻烦,注意到对 $y_i=\ln(x_i/x_0)$而言,不存在约束条件,一旦 y_i 的值确定了,从 y_i 与 $\ln(x_i/x_0)$的关系式和 $\sum_{i=0}^{p}x_i=1$ 就可以直接求出 x_i 的值,因 y_i 可以决定 x_i 的值:

$$\begin{cases} x_i = y_i/\left(1 + \sum_{i=1}^{p} e^{y_i}\right), i = 1,2,\cdots,p, \\ x_0 = 1/\left(1 + \sum_{i=1}^{p} e^{y_i}\right). \end{cases} \tag{5.4}$$

于是,(5.1)式可以改为

$$\begin{cases} u_\alpha = a + \sum_{i=1}^{p} b_\alpha y_{\alpha i} + \varepsilon_\alpha, \alpha = 1,2,\cdots,n, \\ E\varepsilon_\alpha = 0, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \mathrm{Var}(\varepsilon) = \sigma^2 I_n. \end{cases} \tag{5.5}$$

而设计矩阵

$$\underset{n\times(p+1)}{H} = (\mathbb{1}\ y_1 \cdots y_p), y_i = \begin{pmatrix} y_{1i} \\ \vdots \\ y_{ni} \end{pmatrix},$$

$$i = 1,2,\cdots,p$$

y_i 的值可以在$(-\infty,\infty)$内选取.而

$$H'H = \begin{pmatrix} \mathbb{1}' \\ y_1' \\ \vdots \\ y_p' \end{pmatrix} (\mathbb{1}\ y_1 \cdots y_p)$$

$$= \begin{pmatrix} n & y_1'\mathbb{1} & \cdots & y_p'\mathbb{1} \\ \mathbb{1}'y_1 & & & \\ \vdots & & y_i'y_j & \\ \mathbb{1}'y_p & & & \end{pmatrix} = \begin{pmatrix} n & n\bar{y}' \\ n\bar{y} & Y'Y \end{pmatrix},$$

其中

$$\underset{n\times p}{Y} = (y_1 \cdots y_p), \ \bar{y} = \frac{1}{n} Y'\mathbb{1}.$$

这就是如同普通的回归问题一样,对 y 的选择就容易得多.这一点正是加法逻辑正态的方便之处,但这一方法本身显然在回归、试

验设计中对其他分布也是适用的.

§6. 判别分析

由于 x 遵从加法逻辑正态分布时,处理它在变换后的 $y_i = \ln(x_i/x_0), i=1,2,\cdots,p$,可以完全按照正态的情形进行,所以判别分析也就简单了,无论是两总体或多总体的情况,都可沿用正态的结果.在这里我们用判别信息量来导出判别函数.两个密度之间的距离可以有两种度量方法,记 $p_1(x), p_2(x)$ 是两个密度函数,则

$$I(p_1(x), p_2(x)) = \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} \mathrm{d}x, \tag{6.1}$$

$$J(p_1(x), p_2(x)) = I(p_1(x), p_2(x)) + I(p_2(x), p_1(x)), \tag{6.2}$$

都是非负的,是一定意义下度量 $p_1(x)$、$p_2(x)$ 之间差距的量.很明显,$I(p_1(x), p_2(x))$ 对于 $p_1(x)$、$p_2(x)$ 是不对称的,而 $J(p_1(x), p_2(x))$ 是对称的.现在将它用于两个正态总体,并导出相应的表达式.设 $p_1(x)$ 是 $N\left(\mu_1, \underset{p\times p}{\sum\nolimits_1}\right)$, $p_2(x)$ 是 $N\left(\mu_2, \underset{p\times p}{\sum\nolimits_2}\right)$,于是

$$I(p_1(x), p_2(x)) = \int p_1(x) \ln \frac{p_1(x)}{p_2(x)} \mathrm{d}x$$

$$= \frac{1}{2} E\left[\ln \frac{\left|\sum_2\right|}{\left|\sum_1\right|} + (x-\mu_2)' \sum\nolimits_2^{-1} (x-\mu_2) - (x-\mu_1)' \sum\nolimits_1^{-1} (x-\mu_1)\right]$$

$$= \frac{1}{2} \ln \frac{\left|\sum_2\right|}{\left|\sum_1\right|} - \frac{p}{2} + \frac{1}{2} \mathrm{tr} \sum\nolimits_2^{-1} E(x-\mu_2)(x-\mu_2)'.$$

注意此时的求期望 E 是对 $p_1(x)$ 求的,因此

$$E(x-\mu_2)(x-\mu_2)' = \sum_1 + (\mu_1-\mu_2)(\mu_1-\mu_2)'.$$

这样就求得

$$\begin{aligned} & I(p_1(x),p_2(x)) \\ &= \frac{1}{2}\ln\frac{\left|\sum_2\right|}{\left|\sum_1\right|} - \frac{p}{2} \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\sum\nolimits_2^{-1}\sum\nolimits_1 + (\mu_1-\mu_2)(\mu_1-\mu_2)'\right) \\ &= \frac{1}{2}\left[\ln\frac{\left|\sum_2\right|}{\left|\sum_1\right|} + \mathrm{tr}\sum\nolimits_2^{-1}\sum\nolimits_1\right. \\ &\quad \left. + (\mu_1-\mu_2)'\sum\nolimits_2^{-1}(\mu_1-\mu_2) - p\right], \end{aligned} \tag{6.3}$$

同理有

$$\begin{aligned} I(p_2(x),p_1(x)) = \frac{1}{2}&\left[\ln\frac{\left|\sum_1\right|}{\left|\sum_2\right|} + \mathrm{tr}\sum\nolimits_1^{-1}\sum\nolimits_2\right. \\ &\left. + (\mu_1-\mu_2)'\sum\nolimits_1^{-1}(\mu_1-\mu_2) - p\right], \end{aligned} \tag{6.4}$$

于是

$$\begin{aligned} J(p_1(x),p_2(x)) = \frac{1}{2}&\left(\mathrm{tr}\left(\sum\nolimits_1^{-1}\sum\nolimits_2 + \sum\nolimits_2^{-1}\sum\nolimits_1\right)\right. \\ &+ (\mu_1-\mu_2)'\sum\nolimits_1^{-1}(\mu_1-\mu_2) \\ &\left.+ (\mu_1-\mu_2)'\sum\nolimits_2^{-1}(\mu_1-\mu_2)\right) - p. \end{aligned} \tag{6.5}$$

容易看出 $I(p_1(x),p_2(x))$是以 $p_1(x)$为比较基准,看 $p_2(x)$偏

离 $p_1(x)$有多远,同样地,$I(p_2(x),p_1(x))$是以 $p_2(x)$为基准,看 $p_1(x)$偏离 $p_2(x)$有多远.

现在来求出相应的判别函数. 假定 $\mu_1,\mu_2,\Sigma_1,\Sigma_2$ 都是已知的,如果它们是未知的,此时一定有训练样本,从训练样本得到相应估计,因此上述参数用它们的估计量代替,自然也可以认为是已知的. 我们这里处理后一种情况,假定有 n_1 个样本来自 $N(\mu_1,\Sigma_1)$,n_2 个样本来自正态总体 $N(\mu_2,\Sigma_2)$,相应的估计用 $\bar{y}_{(1)}$,$\hat{\Sigma}_{(1)}$和 $\bar{y}_{(2)}$,$\hat{\Sigma}_{(2)}$分别表示,又已知一个新的样本 y,问 y 应归入哪一类,是归入 $N(\mu_1,\Sigma_1)$还是 $N(\mu_2,\Sigma_2)$? 将 y 添加到 $N(\mu_1,\Sigma_1)$的样本中去,得到 μ_1,Σ_1 的新的估计:

$$\hat{\mu}_1=\frac{1}{n_1+1}(n_1\bar{y}_{(1)}+y)$$

$$\begin{aligned}\hat{\Sigma}_1&=\frac{1}{n_1+1}\left(n_1\hat{\Sigma}_{(1)}+(y-\bar{y}_{(1)})(y-\bar{y}_{(1)})'\frac{n_1^2}{n_1+1}\right)\\&=\frac{n_1}{n_1+1}\left(\hat{\Sigma}_{(1)}+\frac{n_1}{n_1+1}(y-\bar{y}_{(1)})(y-\bar{y}_{(1)})'\right).\end{aligned}$$

现在用 $\bar{y}_{(1)}$,$\hat{\Sigma}_{(1)}$代(6.3)中的 μ_1,Σ_1,用 $\hat{\mu}_1$,$\hat{\Sigma}_{(1)}$代(6.3)中 μ_2 和Σ_2,就可求出相应的距离

$$d^2\left(y,N\left(\mu_1,\Sigma_1\right)\right)=\frac{1}{2}\left[\ln\frac{\left|\hat{\Sigma}_1\right|}{\left|\hat{\Sigma}_{(1)}\right|}+\operatorname{tr}\hat{\Sigma}_1^{-1}\hat{\Sigma}_{(1)}\right.$$
$$\left.+(\hat{\mu}_1-\bar{y}_{(1)})'\hat{\Sigma}_1^{-1}(\hat{\mu}_1-\bar{y}_{(1)})-p\right].$$

注意到 $\hat{\mu}_1-\bar{y}_{(1)}=\dfrac{1}{n_1+1}(y-\bar{y}_{(1)})$,又

$$\left|\hat{\Sigma}_1\right|=\left(\frac{n_1}{n_1+1}\right)^p\left|\hat{\Sigma}_{(1)}+\frac{n_1}{n_1+1}(y-\bar{y}_{(1)})(y-\bar{y}_{(1)})'\right|$$

$$= \left| \hat{\Sigma}_{(1)} \right| \left(\frac{n_1}{n_1 + 1} \right)^p \left(1 + \frac{n_1}{n_1 + 1} (y - \bar{y}_{(1)})' \times \hat{\Sigma}_{(1)}^{-1} (y - \bar{y}_{(1)}) \right),$$

因此

$$\left| \hat{\Sigma}_1 \right| / \left| \hat{\Sigma}_{(1)} \right| = \left(\frac{n_1}{n_1 + 1} \right)^p \left(1 + \frac{n_1}{n_1 + 1} (y - \bar{y}_{(1)})' \hat{\Sigma}_{(1)}^{-1} (y - \bar{y}_{(1)}) \right),$$

又

$$\hat{\Sigma}_1^{-1} = \frac{n_1 + 1}{n_1} \left(\hat{\Sigma}_{(1)} + \frac{n_1}{n_1 + 1} (y - \bar{y}_{(1)})(y - \bar{y}_{(1)})' \right)^{-1}$$

$$= \frac{n_1 + 1}{n_1} \left(\hat{\Sigma}_{(1)}^{-1} + \hat{\Sigma}_{(1)}^{-1} (y - \bar{y}_{(1)}) v^{-1} (y - \bar{y}_{(1)})' \times \hat{\Sigma}_{(1)}^{-1} \frac{n_1}{n_1 + 1} \right),$$

其中

$$v = \left(1 + \frac{n_1}{n_1 + 1} (y - \bar{y}_{(1)})' \hat{\Sigma}_{(1)}^{-1} (y - \bar{y}_{(1)}) \right),$$

因此

$$\mathrm{tr} \hat{\Sigma}_1^{-1} \hat{\Sigma}_{(1)} = \frac{n_1 + 1}{n_1} (p + v^{-1}(v - 1)) = \frac{n_1 + 1}{n_1} (p + 1 - v^{-1}).$$

于是

$$\mathrm{d}^2 \left(y, N\left((\mu_1, \Sigma_1 \right) \right) = \frac{1}{2} \left[p \ln \frac{n_1}{n_1 + 1} + \ln v + \frac{n_1 + 1}{n_1} (p + 1 - v^{-1}) + \frac{n_1^2}{(n_1 + 1)^2} (y - \bar{y}_{(1)})' \hat{\Sigma}_1^{-1} (y - \bar{y}_{(1)}) - p \right]$$

$$
= \frac{1}{2}\Big[p\ln \frac{n_1}{n_1+1} + \ln v + \frac{n_1+1}{n_1}(p+1-v^{-1})
$$

$$
+ \frac{n_1}{n_1+1}(y-\bar{y}_{(1)})'\sum\nolimits_{(1)}^{-1}(y-\bar{y}_{(1)}) + \frac{1}{v}(v-1)^2 - p\Big]
$$

$$
= \frac{1}{2}\Big[p\ln \frac{n_1}{n_1+1} + \ln v + \frac{n_1+1}{n_1}(p+1-v^{-1})
$$

$$
+ v - 1 + (v-1)^2/v - p\Big].
$$

当 n_1 相当大时，$n_1/(n_1+1)$ 可以看成是 1，上式就简化为

$$
d^2\left(y, N\left(\mu_1, \sum\nolimits_1\right)\right) = v - 1 = (y-\bar{y}_{(1)})'\sum\nolimits_{(1)}^{-1}(y-\bar{y}_{(1)}),
$$

它实际上就是 y 对 $N(\mu_1,\Sigma_1)$ 的马哈拉诺比斯距离，这一结论与用贝叶斯方法，费歇方法导出的是一致的. 当我们已知 μ_i,Σ_i 时，可以认为 n_1,n_2 都是 ∞，因此 $\frac{n_1}{n_1+1}$ 用 1 代是完全正确的，这样对上述结论就会有更清晰的解释，它确实反映了观察点 y 与总体 $N(\mu_1,\Sigma_1)$ 的距离，对于 $N(\mu_2,\Sigma_2)$，自然也是得到相似的结论，这就不用重复了.

上面我们处理的是两总体的判别，对于多总体的判别，自然可以直接推广，只须计算观察点 y 与各总体的距离就可以判断它该归入哪一类，所以判别分析的问题也就得到了解决.

习 题 三

1. 利用表 1.1 的数据，去算一下有关的估计(可以仿照例中的方法，将有些合并处理).

2. 证明公式(3.4)：$z = F^+ y, y = Fz$.

(F^+ 的定义：若对给定的一个矩阵 F，存在 X 满足：

$$
FXF = F, XFX = X, (FX)' = FX, (XF)' = XF,
$$

则算 X 是 F^+. 可以证明 F^+ 存在而且唯一，且有

$$
F^+ = F'(FF')^+ = (F'F)^+ F').
$$

3. 计算医院资料(表 1.1)相应的均值、方差协方差矩阵.要比较两个医院的疗效,应如何进行? 如果直接计算成分 x 来比较,利用 y 来比较,对 $\bar{x}$ 考虑到三年共有 36 个月,认为可以用大样本近似,当作正态来处理;而对 y 用加法逻辑正态来处理.这两种比较结果在结论中有无差别?

4. 用医院的资料(表 1.1),考虑两总体的判别问题.

5. 对试验设计.如果考虑 $H'H = aI$,a 是一个指定的常数,这时导出的设计阵是"正交"设计.就成分向量只有三个分量,用(5.1)模型,求出的设计是什么? 用模型(5.5)求出设计阵是什么? 两者有什么区别?

6. 对于混料试验设计,添加若干个试验的设计可以仿 §5 的讨论进行,求出相应的 A 最优的设计应满足的方程.

参 考 文 献

[1] 张尧庭、方开泰(1982,1998),多元统计分析引论,科学出版社.

[2] Aitchison, J. (1986), The Statistical Analysis of Compositional Data, Chapman & Hall. (中译本:周蒂等译,成分数据的统计分析(1989),中国地质大学出版社)

第四章　狄氏分布的统计分析

§1. 准备知识

狄氏分布的密度函数并不复杂，但是它的统计分析和逻辑正态不同.逻辑正态是将成分向量 x 作适当的变换后，变成正态分布，因此处理变换后的变量就和正态相同.而狄氏分布没有这一方便，就是参数估计，极大似然估计也给不出明确的表达式，只能用数值的方法求解，因此无论是估计和检验都有它的困难.

近年来，广义矩估计法的影响越来越明显，对于狄氏分布的参数用广义矩估计来获得有关参数的估计是完全可行的，因此在这一节我们扼要地介绍广义矩估计法的方法，所得估计的性质，以及有关的一些结论.

通常的矩估计法的基础是两点：

(a)总体的矩是总体参数的函数.如用 ξ 表示总体分布相应的特征值，总体参数是期望值 μ 和总体方差 σ^2，于是就有

$$\begin{cases} E\xi = \mu, \\ E\xi^2 = \sigma^2 + \mu^2. \end{cases} \tag{1.1}$$

因此 ξ 的各阶矩都与总体参数有联系，它们可以用参数的函数表示出来.

(b)用样本的矩来代替总体的矩，从关系式中解出参数，就获得参数的估计量.在(1.1)中分别用样本 $x_1, \cdots, x_n$ 的均值 $\bar{x}$ 与二阶矩 $\frac{1}{n}\sum\limits_{\alpha=1}^{n} x_\alpha^2$ 代 $E\xi$ 和 $E\xi^2$，就可解出 μ 与 σ^2 的估计

$$\hat{\mu} = \bar{x}, \hat{\sigma}^2 = \frac{1}{n}\sum_{\alpha=1}^{n} x_\alpha^2 - \bar{x}^2 = \frac{1}{n}\sum_{\alpha=1}^{n}(x_\alpha - \bar{x})^2.$$

从(a)、(b)不难看出，为什么只限于总体的矩呢？为什么不可

以更广泛地考虑各种与参数有关的 ξ 的函数呢？如果我们用 θ 表示要估计的参数，θ 可以是向量．上述(a)、(b)两条可以换成

(a)寻找合适的函数 $h_i(\xi,\theta)$，使得

$$Eh_i(\xi;\theta)=0,\qquad i=1,2,\cdots,m,$$

(1.1)中，$h_1(\xi;\theta)=\xi-\mu$，$h_2(\xi;\theta)=\xi^2-\sigma^2-\mu^2$．

(b)用样本 x_α 代替 $h_i(\xi;\theta)$ 中的 ξ，用对样本代人后的 $h_i(x_\alpha;\theta)$求平均，代替取期望值，即从

$$\frac{1}{n}\sum_{\alpha=1}^{n}h_i(x_\alpha;\theta)=0,i=1,2,\cdots,m,\tag{1.2}$$

解出 θ，从而获得 θ 的估计量．如果 θ 中有 k 个参数，当 $m=k$ 时可以求解，当 $m>k$ 时，就要考虑如何去求合理的估计量$\hat{\theta}$．

如果令

$$\begin{cases}g_i(x;\theta)=\dfrac{1}{n}\displaystyle\sum_{\alpha=1}^{n}h_i(x_\alpha;\theta),\\ g(x;\theta)=(g_1(x;\theta),g_2(x;\theta),\cdots,g_m(x;\theta))',\end{cases}\tag{1.3}$$

那么 $g(x;\theta)$就是一个向量，要从 $g(x;\theta)$越接近于0越好来考虑，就是选估计$\hat{\theta}$使$g(x;\theta)$的向量长度 $\|g(x;\theta)\|$ 越小越好．也即估计量$\hat{\theta}$使 $\|g(x;\theta)\|$ 达到最小，用式子表示，即$\hat{\theta}$满足

$$\|g(x;\hat{\theta})\|=\min_{\theta}\|g(x;\theta)\|.$$

然而，向量的长度可以有种种定义的方法，怎样定义使解出的$\hat{\theta}$具有好的性质，这些都是从理论、方法上应该讨论的．

目前对于这一方法的理论性的探讨有一些结果，目的是选一个合适的度量，使得采用迭代方法后，在比较宽松的条件下，$\hat{\theta}$可以收敛于θ 的真值．有关这些可以参看本章的参考文献[1]．

§2. 估　　计

对于狄氏分布，同样有估计和检验的问题．我们沿用上一章的

记号，主要区别在于不加声明时，总假定本章的样本来自狄氏分布.

$x_{(1)},\cdots,x_{(n)}$都是 $p+1$ 维的向量，它们组成的样本矩阵

$$\underset{n\times(p+1)}{X}=\begin{pmatrix}x'_{(1)}\\ \vdots\\ x'_{(n)}\end{pmatrix}=(x_{\alpha i}),$$

$$\alpha=1,2,\cdots,n,i=0,1,\cdots,p.$$

相应的样本均值向量 $\bar{x}$ 和方差协差矩阵S 是

$$\bar{x}=\frac{1}{n}X'\mathbb{1},S=\frac{1}{n-1}L_{xx}$$

$$=\frac{1}{n-1}X'(I_n-\frac{1}{n}\mathbb{1}\mathbb{1}')X.$$

假定总体的分布密度是

$$\Gamma\left(\sum_{i=0}^{p}a_i\right)\prod_{i=0}^{p}\frac{x_i^{a_i-1}}{\Gamma(a_i)},$$

$$a_i>0,x_i>0,i=0,1,\cdots,p,\sum_{i=0}^{p}x_i=1.$$

从第一章就知道成分向量 $x=(x_0,x_1,\cdots,x_p)'$的一、二阶矩是

$$\begin{cases}Ex_i=\dfrac{a_i}{A},\\ \mathrm{Var}(x_i)=a_i(A-a_i)/[A^2(A+1)],\\ \qquad i,j=0,1,\cdots,p,\\ \mathrm{Cov}(x_i,x_j)=-a_ia_j/[A^2(A+1)]i\neq j,\end{cases}\tag{2.1}$$

其中 $A=\sum_{i=0}^{p}a_i$.因此一共只有 $p+1$ 个参数.怎样估计这 $p+1$ 个参数呢?

(i)用一般的矩估计法

用样本均值估计总体均值，样本协差阵估计总体的协差阵，这时就有

$$
\begin{cases}
\left(\frac{\widehat{a_i}}{A}\right) = \bar{x}_i = \frac{1}{n}\sum_{\alpha=1}^{n} x_{\alpha i}, \quad i = 0,1,\cdots,p, \\
\left(\widehat{\frac{a_i(A-a_i)}{A^2(A+1)}}\right) = S_{ii} = \frac{1}{n-1}\sum_{\alpha=1}^{n}(x_{\alpha i} - \bar{x}_i)(x_{\alpha i} - \bar{x}_i), \\
\left(\widehat{\frac{-a_i a_j}{A^2(A+1)}}\right) = S_{ij} = \frac{1}{n-1}\sum_{\alpha=1}^{n}(x_{\alpha i} - \bar{x}_i)(x_{\alpha j} - \bar{x}_j), \\
i \neq j, i, j = 0,1,\cdots,p.
\end{cases}
\tag{2.2}
$$

(2.2)给出的关系式有$(p+1)(p+4)/2$个，大大地多于要估计的$p+1$个.可见这时简单地用一般的矩估计法就会遇到困难，它必然要引向广义矩估计法.

(ii)用最大似然估计法

样本相应的联合密度是

$$
\left[\frac{\Gamma(A)}{\prod_{i=0}^{p}\Gamma(a_i)}\right]^n \prod_{\alpha=1}^{n}\prod_{i=0}^{p} x_{\alpha i}^{a_i-1}.
$$

用 L 表示似然函数，则

$$
\begin{aligned}
l = \ln L = & n\ln\Gamma(A) - \sum_{i=0}^{p} n\ln\Gamma(a_i) \\
& + \sum_{i=0}^{p}(a_i - 1)\left(\sum_{\alpha=1}^{n}\ln x_{\alpha i}\right).
\end{aligned}
$$

对 a_i 求偏导数，让这些偏导数为 0，就得方程

$$
0 = \frac{\partial l}{\partial a_i} = \frac{n\Gamma'(A)}{\Gamma(A)} - \frac{n\Gamma'(a_i)}{\Gamma(a_i)} + \sum_{\alpha=1}^{n}\ln x_{\alpha i},
$$

$$
i = 0,1,\cdots,p, \tag{2.3}
$$

其中 $\Gamma'(a)$是 $\Gamma(a)$对 a 的导数.方程(2.3)是很难直接求解的，只能用数值的解法.

如果利用$\frac{\Gamma'(z)}{\Gamma(z)}$的展开式

$$
\frac{\Gamma'(z)}{\Gamma(z)} = -C - \frac{1}{z} + \sum_{k=1}^{\infty}\left(\frac{1}{k} - \frac{1}{z+k}\right),
$$

略去高阶的项，则从(2.3)可得方程

$$\frac{1}{a_i}-\frac{1}{A}=-\frac{1}{n}\sum_{a=1}^{n}\ln x_{ai},\qquad i=0,1,\cdots,p,$$

也即

$$a_i=\frac{1}{\frac{1}{A}-\ln g_i}=\frac{A}{1-A\ln g_i},i=0,1,\cdots,p,$$

其中

$$g_i=\left(\prod_{a=1}^{n}x_{ai}\right)^{\frac{1}{n}},i=0,1,\cdots,p.$$

对上式两端求和,得

$$\sum_{i=0}^{p}a_i=A=A\sum_{i=0}^{p}\frac{1}{1-A\ln g_i},$$

因此得估计

$$\begin{cases}\hat{A}\text{ 是方程}\quad 1=\sum_{i=0}^{p}\frac{1}{1-A\ln g_i}\text{ 的根},\\ \hat{a}_i=\hat{A}/(1-\hat{A}\ln g_i),i=0,1,\cdots,p.\end{cases}\tag{2.4}$$

(2.4)式中 $\hat{A}$ 满足的方程有不少根,它是一个 $p+1$ 阶的多项式,因而有 $p+1$ 个根,这样还是不能得到一个显式的估计.

值得注意的是,a_i 的估计从(2.4)式可以看出,它们都是样本 x_{ai} 的几何平均值 $\left(\prod_{a=1}^{n}x_{ai}\right)^{\frac{1}{n}}$ 的函数.

(iii)用广义矩估计法

广义矩估计法,要求寻找一些函数 $h(x;\theta)$ 使 $Eh(x;\theta)=0$ 对一切 θ 都成立.最自然的方法是从成分向量 x 的联合密度出发,设 x 的联合密度为 $f(x;\theta)$,于是有

$$\int f(x;\theta)\mathrm{d}x=1,\forall\theta\text{ 成立}.$$

两边对 θ 求微商,于是有

$$0=\int\frac{\partial f}{\partial\theta}\mathrm{d}x=\int\left(\frac{\partial f}{\partial\theta}\right)\frac{1}{f}\cdot f(x;\theta)\mathrm{d}x$$

$$= \int \left(\frac{\partial \ln f}{\partial \theta}\right) f(x;\theta) \mathrm{d}x$$

$$= E \frac{\partial \ln f}{\partial \theta},$$

可见取 $h = \frac{\partial \ln f}{\partial \theta}$ 是合适的，它得到的 h 就是似然方程(2.3)右端所相应的函数，因为 $l = \ln f, \frac{\partial l}{\partial \theta} = \frac{\partial \ln f}{\partial \theta} = h$. 从这里更可以看出广义矩估计与最大似然估计的关系. 最大似然估计就是选择了一组特殊的 h_i，并且 h_i 的个数与参数的数目正好相当. 下面对狄氏分布，我们来验证一下这个看法.

成分向量 $x = (x_0, x_1, \cdots, x_p)'$ 的联合密度在 D_p 上的积分是 1，即

$$\int_{D_p} \frac{\Gamma(A)}{\prod_{i=0}^{p} \Gamma(a_i)} \left(\prod_{i=0}^{p} x_i^{a_i - 1}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_p = 1,$$

也即

$$\int_{D_p} \left(\prod_{i=0}^{p} x_i^{a_i - 1}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_p = \frac{\prod_{i=0}^{p} \Gamma(a_i)}{\Gamma(A)}.$$

两边对 a_i 求偏微商，得：对 $i = 0, 1, \cdots, p$ 有

$$\int_{D_p} (\ln x_i) \left(\prod_{i=0}^{p} x_i^{a_i - 1}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_p$$

$$= \frac{\prod_{i=0}^{p} \Gamma(a_i)}{[\Gamma(A)]^2} \left(\frac{\Gamma'(a_i)}{\Gamma(a_i)} \Gamma(A) - \Gamma'(A)\right),$$

也即

$$\frac{\Gamma(A)}{\prod_{i=0}^{p} \Gamma(a_i)} \int_{D_p} (\ln x_i) \left(\prod_{i=0}^{p} x_i^{a_i - 1}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_p = \frac{\Gamma'(a_i)}{\Gamma(a_i)} - \frac{\Gamma'(A)}{\Gamma(A)}.$$

上式左端就是 $E \ln x_i$，因此上式就是

$$E\left(\ln x_i + \frac{\Gamma'(A)}{\Gamma(A)} - \frac{\Gamma'(a_i)}{\Gamma(a_i)}\right) = 0, i = 0,1,\cdots,p.$$

因此取 $h_i(x;\theta) = \ln x_i + \dfrac{\Gamma'(A)}{\Gamma(A)} - \dfrac{\Gamma'(a_i)}{\Gamma(a_i)}, i = 0,1,\cdots,p$ 后

$$\begin{aligned}
&\frac{1}{n}\sum_{\alpha=1}^{n} h_i(x_{(\alpha)};\theta) \\
&= \frac{1}{n}\sum_{\alpha=1}^{n} h(x_{\alpha i};\theta) = \frac{1}{n}\sum_{\alpha=1}^{n}\left(\ln x_{\alpha i} + \frac{\Gamma'(A)}{\Gamma(A)} - \frac{\Gamma'(a_i)}{\Gamma(a_i)}\right) \\
&= \frac{1}{n}\left[\sum_{\alpha=1}^{n}\ln x_{\alpha i} + \frac{n\Gamma'(A)}{\Gamma(A)} - \frac{n\Gamma'(a_i)}{\Gamma(a_i)}\right].
\end{aligned}$$

很明显,上式右端就是似然方程(2.3)相应的函数.

(iv)采用线性模型来估计

我们还是从$(x_0, x_1, \cdots, x_p)' = x$ 这个成分向量的 n 个样本 $x_{(1)}, \cdots, x_{(n)}$形成的矩阵 X 开始,此时

$$\underset{n\times(p+1)}{X} = (x_{\alpha i}) = \begin{pmatrix} x'_{(1)} \\ \vdots \\ x'_{(n)} \end{pmatrix} = (\boldsymbol{x}_0 \quad \boldsymbol{x}_1 \quad \cdots \quad \boldsymbol{x}_p),$$

相应的样本均值 $\bar{x}$ 与协方差矩阵S 是

$$\bar{x} = \frac{1}{n}X'\mathbb{1}, S = \frac{1}{n-1}L_{xx},$$

其中

$$L_{xx} = X'(I_n - \frac{1}{n}\mathbb{1}\mathbb{1}')X.$$

为了估计时表达式更简明,将狄氏分布的参数 $a_0, a_1, \cdots, a_p$ 作一变换,令

$$A = \sum_{i=0}^{p} a_i, \qquad b_i = a_i/A, i = 0,1,\cdots,p$$

于是可将(2.1)式改写为

$$
\begin{cases}
Ex_i = b_i, \\
\mathrm{Var}(x_i) = \dfrac{1}{A+1} b_i(1-b_i), \\
\mathrm{Cov}(x_i, x_j) = -\dfrac{1}{A+1} b_i b_j, \quad i \neq j, i, j = 0, 1, \cdots, p,
\end{cases}
\tag{2.5}
$$

此时共有 $p+1$ 个参数,因为 $\sum_{j=0}^{p} b_j = 1$. 用向量和矩阵来表示,就可以导出如下的形式:

记

$$
D(b) = \begin{pmatrix} b_0 & & & 0 \\ & b_1 & & \\ & & \ddots & \\ 0 & & & b_p \end{pmatrix}, b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix},
$$

于是有

$$
Ex = b, \mathrm{Var}(x) = \frac{1}{A+1}(D(b) - bb'). \tag{2.6}
$$

很明显,参数 b 与期望值有关,参数 A 与 x 的协方差矩阵有关, $\mathrm{Var}(x)$是退化的,因为

$$
(D(b) - bb')\mathbb{1} = b - b = 0.
$$

尽管 $D(b) - bb'$ 是退化的,但它的"+"逆是唯一确定的. 不难证明(留作习题去求出这个表达式)

$$
\begin{aligned}
(D(b) - bb')^{+} = D(b^{-1}) &- \frac{1}{h(p+1)} \mathbb{1}\mathbb{1}' \\
&- \frac{1}{p+1}[b^{-1}\mathbb{1}' + \mathbb{1}(b^{-1})'],
\end{aligned}
\tag{2.7}
$$

其中

$$
b^{-1} = \begin{pmatrix} b_0^{-1} \\ \vdots \\ b_p^{-1} \end{pmatrix}, h = \left(\frac{1}{p+1} \sum_{i=0}^{p} b_i^{-1}\right)^{-1}.
$$

记 $A=D(b)-bb'$,要证明(2.7)式是 A^+,只须证明 AA^+,A^+A 是对称的,$AA^+A=A$,$A^+AA^+=A^+$,实际上可以算得

$$AA^+=I-\frac{1}{p+1}\mathbb{1}\mathbb{1}',$$

$$A^+A=I-\frac{1}{p+1}\mathbb{1}\mathbb{1}',$$

因此自然就有

$$AA^+A=A,A^+AA^+=A^+.$$

难点是在于如何找到这一表达式.(提示见习题)

有了这些准备,就可以用两种不同的方式将它转化为线性模型的问题.

一种是考察每一个成分向量的样本 $\boldsymbol{x}_i$,$i=0,1,\cdots,p$,易见有

$$\boldsymbol{x}_i'=(x_{1i},\cdots,x_{ni}),$$

$$\begin{cases}E\boldsymbol{x}_i=\underset{n\times1}{\mathbb{1}}b_i,\\ \mathrm{Var}(\boldsymbol{x}_i)=\dfrac{1}{\mathrm{A}+1}b_i(1-b_i)I_n,\end{cases}\tag{2.8}$$

对 $i=0,1,\cdots,p$ 都成立.利用均值$\frac{1}{n}\boldsymbol{x}_i'\mathbb{1}=\bar{x}_i$ 估计 b_i,利用方差

$$S_i^2=\frac{1}{n-1}\boldsymbol{x}_i'\left(I_n-\frac{1}{n}\mathbb{1}\mathbb{1}'\right)\boldsymbol{x}_i$$

去估计 $b_i(1-b_i)/(A+1)$,于是就有

$$\begin{cases}\hat{b}_i=\bar{x}_i\\ \hat{\mathrm{A}}_i=\dfrac{\bar{x}_i(1-\bar{x}_i)}{S_i^2}-1\end{cases}\tag{2.9}$$

对 $i=0,1,\cdots,p$ 都成立.这样我们就有 $p+1$ 个 A 的估计量$\hat{A}$,为了区别我们用 $\hat{\mathrm{A}}_i$ 表示用第i 个成分向量样本 $\boldsymbol{x}_i$ 导出的A 的估计,这就是(2.9)式中的 $\hat{A}_i$.

由(2.9)式给出的 $\hat{A}$能否保证是一个正数呢?下面的引理就回答了这个问题.

引理 2.1 设 $0<u_i<1, i=1,2,\cdots,n$，则

$$\sum_{i=1}^{n}(u_i-\bar{u})^2 \leqslant n\bar{u}(1-\bar{u}).$$

证明 $\sum_{i=1}^{n}(u_i-\bar{u})^2=\sum_{i=1}^{n}u_i^2-n\bar{u}^2$，由于 $0<u_0<1, i=1,2,\cdots,n$，因此 $u_i^2<u_i$. 于是有

$$\sum_{i=1}^{n}(u_i-\bar{u})^2 \leqslant \sum_{i=1}^{n}u_i-n\bar{u}^2=n\bar{u}(1-\bar{u}).$$

从引理 2.1 可以看出(2.9)式中的 $\hat{A}_i>0$ 通常是不会有问题的. 但这些 $\hat{A}_i$ 是不独立的，如何从这些 $\hat{A}_i$ 去综合一个更好的 $\hat{A}$，这是应讨论的，如果我们采用另一个方式，就可以直接得到一个 $\hat{A}$，这就是下面的第二种方式.

我们直接从成分向量的样本矩阵 X 开始，今有

$$E\underset{n\times(p+1)}{X}=\underset{n\times 1}{\mathbb{1}}\underset{1\times(p+1)}{b'}, \tag{2.10}$$

X 的各行独立、同协差阵 $\frac{1}{A+1}(D(b)-bb')$.

要注意此时 $b'\mathbb{1}=1$ 是一个约束条件，因此(2.10)这个多元线性模型是带约束条件的线性模型，而且协方差矩阵是退化的.

现在将约束条件的解全部表示出来，把模型(2.10)化为无约束的模型，然后直接用多元线性模型的结果来求解. 当然也可以去掉 X 中的一列，变成无约束的情形. 这两种方法所得的结果是一样的，前一种可以参见文献[2]，我们这里采用第二种，去掉 x_0 对应的样本，就得到

$$X_*=(\boldsymbol{x}_1\,\boldsymbol{x}_2\cdots\boldsymbol{x}_p),$$

于是令 $b_*=(b_1,b_2,\cdots,b_p)'$ 后，就有

$$E\underset{n\times p}{X_*}=\mathbb{1}b'_*, \tag{2.11}$$

X_* 的各行独立，同协方差阵 $\frac{1}{A+1}(D(b_*)-b_*b'_*)$.

此时，直接用多元线性模型的结果，就得到

$$\hat{b}_* = \frac{1}{n} X'_* \mathbb{1}, \tag{2.12}$$

$$(n-p)\hat{\sum} = X'_* \left(I_n - \frac{1}{n} \mathbb{1}\mathbb{1}' \right) X_*, \tag{2.13}$$

其中

$$\sum = \frac{1}{A+1}(D(b_*) - b_* b'_*).$$

利用(2.12)和(2.13)就可以求出 A 的估计量 $\hat{A}$. 由于$\sum$中只有一个参数 A 需要估计,而 b_* 用 $\hat{b}_*$ 估计是无偏的最小二乘估计,具有优良性.从(2.13)式可以导出 $\hat{A}$的几个不同的估计.

利用 $\mathrm{tr}\sum$,由于

$$\begin{aligned} \mathrm{tr}\sum &= \frac{1}{A+1} \mathrm{tr}(D(b_*) - b_* b'_*) \\ &= \frac{1}{A+1}\left(\sum_{i=1}^{p} b_i(1-b_i) \right), \end{aligned}$$

于是有

$$\hat{A} = \frac{\sum_{i=1}^{p} \hat{b}_i(1-\hat{b}_i)}{\mathrm{tr}\hat{\sum}} - 1. \tag{2.14}$$

利用$\mathbb{1}'\sum\mathbb{1}$,由于

$$\begin{aligned} \mathbb{1}'\sum\mathbb{1} &= \frac{1}{A+1}(\mathbb{1}'D(b_*)\mathbb{1} - \mathbb{1}'b_* b'_* \mathbb{1}) \\ &= \frac{1}{A+1}(1 - b_0 - (1-b_0)^2) \\ &= \frac{1}{A+1} b_0(1-b_0). \end{aligned}$$

因此有

$$\hat{A} = \frac{\hat{b}_0(1-\hat{b}_0)}{\mathbb{1}'\hat{\sum}\mathbb{1}} - 1. \tag{2.15}$$

利用$\sum$的行列式，由于

$$\left|\sum\right| = \left(\frac{1}{A+1}\right)^p \left|D(b_*) - b_* b'_*\right|$$

$$= \left(\frac{1}{A+1}\right)^p (1 - b'_* D^{-1}(b_*) b_*) \left|D(b_*)\right|$$

$$= \left(\frac{1}{A+1}\right)^p (1 - b'_* \mathbb{1}) \left|D(b_*)\right|$$

$$= \left(\frac{1}{A+1}\right)^p \prod_{j=0}^{p} b_j$$

因此有

$$\hat{A} = \left(\frac{\prod_{j=0}^{p} \hat{b}_j}{\left|\hat{\sum}\right|}\right)^{\frac{1}{p}} - 1. \tag{2.16}$$

很明显(2.14)，(2.15)，(2.16)给出的 $\hat{A}$ 是不同的. 不难看出，利用$\sum$的函数，还可以给出不同的估计. 为了区别这三个估计，我们用 $\hat{A}(1)$、$\hat{A}(2)$、$\hat{A}(3)$依次表示(2.14)、(2.15)、(2.16)的估计量. 注意(2.9)式的估计量为 $\hat{A}_i, i=0,1,\cdots,p$.

为了比较这些 A 的估计量的好坏，比较它们各自渐近分布的方差，是一种常用的方法，这些估计量都不是无偏的，但是它们的偏差在大样本时可以忽略，所以只比较渐近方差的大小，此时要用参考文献[3]中的一条定理，我们不加证明来引用，作为本节的引理 2.2.

引理 2.2 设 $x_1, x_2, \cdots, x_n, \cdots$是独立同分布的随机向量序列，$x_i$ 均为 $p\times 1$ 的向量，有

$$Ex_i = \theta, \mathrm{Var}(x_i) = \Omega.$$

又 $f(t)$是 $p\times 1$ 向量 t 的函数，具有一、二、三阶各种偏微商，且均为 t 的连续函数，取

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \beta = \left.\frac{\partial f}{\partial t}\right|_{t=\theta},$$

则当 $\beta'\Omega\beta\neq 0$ 时，

$$\sqrt{n}(f(\bar{x}_n)-f(\theta))$$

的极限分布是 $N(0,\beta'\Omega\beta)$.

引理 2.2 告诉我们，样本均值 $\bar{x}_n$ 的函数 $f(\bar{x}_n)$ 的渐近分布是 $N\left(f(\theta),\frac{1}{n}\beta'\Omega\beta\right)$，渐近方差就是 $\beta'\Omega\beta/n$. 于是只需比较 $\beta'\Omega\beta$ 的值，就可以判断谁好谁差. 关键是将相应于引理 2.2 中的 x_i，$f(\cdot)$，β，Ω 都找到. 我们先讨论 $\hat{A}_i$，然后讨论 $\hat{A}(i)$，$i=1,2,3$.

注意到 $\hat{A}_i$ 相应的 f 都是相同的，只是各自用成分第 i 个分量的均值与方差来作为自变量，所以只须讨论其中的一个，其余均可类似地求得. 由于 $\hat{b}_i$ 本身就是样本的均值，而

$$S_i^2=\left(\sum_{\alpha=1}^{n}x_{\alpha i}^2-n\bar{x}_i^2\right)\Big/(n-1)$$
$$=\frac{n}{n-1}\left[\frac{1}{n}\sum_{\alpha=1}^{n}x_{\alpha i}^2-\bar{x}_i^2\right].$$

考虑 i 固定，

$$y_\alpha=\begin{pmatrix}x_{\alpha i}\\x_{\alpha i}^2\end{pmatrix}\triangleq\begin{pmatrix}y_{\alpha 1}\\y_{\alpha 2}\end{pmatrix},\alpha=1,2,\cdots,n,\cdots$$

$$f(\bar{y}_1,\bar{y}_2)=\frac{\bar{y}_1(1-\bar{y}_1)}{\bar{y}_2-\bar{y}_1^2},$$

其中

$$\bar{y}_1=\frac{1}{n}\sum_{\alpha=1}^{n}y_{\alpha 1}=\frac{1}{n}\sum_{\alpha=1}^{n}x_{\alpha i}=\hat{b}_i,$$

$$\bar{y}_2=\frac{1}{n}\sum_{\alpha=1}^{n}y_{\alpha 2}=\frac{1}{n}\sum_{\alpha=1}^{n}x_{\alpha i}^2.$$

于是

$$S_i^2=\frac{n}{n-1}(\bar{y}_2-\bar{y}_1^2).$$

当 n 很大时，可以认为 $S_i^2=\bar{y}_2-\bar{y}_1^2$，于是就可以用引理 2.2 求出 $\hat{A}_i$ 的渐近分布. 今

$$\theta = \mathrm{E}\begin{pmatrix} x_i \\ x_i^2 \end{pmatrix} = \begin{pmatrix} b_i \\ \dfrac{b_i}{A+1}(1 + Ab_i) \end{pmatrix},$$

$$\Omega = \begin{pmatrix} \mathrm{Var}(x_i) & \mathrm{Cov}(x_i, x_i^2) \\ \mathrm{Cov}(x_i^2, x_i) & \mathrm{Var}(x_i^2) \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{b_i(1 - b_i)}{A + 1} & \dfrac{2b_i(1 - b_i)(1 + Ab_i)}{(A + 1)(A + 2)} \\ \dfrac{2b_i(1 - b_i)(1 + Ab_i)}{(A + 1)(A + 2)} & \dfrac{b_i(1 + b_i)A\left[b(A + 1) + (4A^2 - 6)b_i - (4A + 6)Ab_i^2\right]}{(A + 1)^2(A + 2)(A + 3)} \end{pmatrix}.$$

令

$$f(t_1, t_2) = \frac{t_1 - t_1^2}{t_2 - t_1^2},$$

因此

$$\frac{\partial f}{\partial t_1} = \frac{t_1^2 - 2t_1t_2 + t_2}{(t_2 - t_1^2)^2},$$

$$\frac{\partial f}{\partial t_2} = -\frac{t_1(1 - t_1)}{(t_2 - t_1^2)^2}.$$

在上两式中,用 b_i 代 t_1,$\dfrac{b_i(1 + Ab_i)}{A + 1}$ 代 t_2,就可以求出 β 的值,即

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \frac{A + 1}{b_i(1 - b_i)}\begin{pmatrix} 1 + 2Ab_i \\ -(A + 1) \end{pmatrix},$$

因此

$$\beta'\Omega\beta = \frac{(A + 1)^2}{b_i^2(1 - b_i)^2}\Big[\frac{b_i(1 - b_i)(1 + 2Ab_i)^2}{A + 1}$$

$$- \partial\,\frac{2b_i(1 - b_i)(1 + Ab_i)}{(A + 1)(A + 2)}(A + 1)(1 + 2Ab_i)$$

$$+\frac{b_i(1+b_i)A[6(A+1)+(4A^2-6)b_i-(4A+6)Ab_i^2]}{(A+2)(A+3)}\Bigg]$$

$$=\frac{(A+1)^2}{b_i(1-b_i)^2}[(6A^3+9A^2-5A-6)$$

$$+(4A^4+2A^3-14A^2-A+6)b_i$$

$$-A(4A^3+10A^2-16A-6)b_i^2+A^2(2A-3)b_i^3].$$

完全同样的方法,可以求出 $\hat{A}(1)$、$\hat{A}(2)$、$\hat{A}(3)$的渐近方差.

$\hat{A}(1)$、$\hat{A}(2)$的表达式相似,$\hat{A}(3)$很不相同,我们详细推导$\hat{A}(1)$渐近方差表达式,将 $\hat{A}(2)$和 $\hat{A}(3)$留作练习.

注意到

$$\hat{A}(1)=\frac{\sum_{i=1}^{p}\hat{b}_i(1-\hat{b}_i)}{\operatorname{tr}\hat{\Sigma}}-1$$

$$=\frac{\sum_{i=1}^{p}\hat{b}_i(1-\hat{b}_i)}{\sum_{i=1}^{p}S_i^2}-1,$$

因此相应的 f 是 $2p$ 个变量的函数,即可取

$$f(t_1,t_2,\cdots,t_p;r_1,\cdots,r_p)=\frac{\sum_{i=1}^{p}t_i(1-t_i)}{\sum_{i=1}^{p}(r_i-t_i^2)},$$

于是

$$\frac{\partial f}{\partial t_i}=\frac{t_i^2-2t_ir_i+r_i}{\left[\sum_{i=1}^{p}(r_i-t_i^2)\right]^2},i=1,2,\cdots,p,$$

$$\frac{\partial f}{\partial r_i} = -\frac{\sum_{i=1}^{p} t_i(1 - t_i)}{\left[\sum_{i=1}^{p}(r_i - t_i^2)\right]^2}, i = 1,2,\cdots,p.$$

因此可以求出 $\beta = \begin{pmatrix} \beta_{(1)} \\ \beta_{(2)} \end{pmatrix}, \beta_{(i)} = \begin{pmatrix} \beta_{i1} \\ \vdots \\ \beta_{ip} \end{pmatrix}, i = 1,2$,其中 β_{ij} 为

$$\beta_{1j} = \left.\frac{\partial f}{\partial t_i}\right|_{\mu}, \beta_{2j} = \left.\frac{\partial f}{\partial r_i}\right|_{\mu}, j = 1,2,\cdots,p.$$

注意到

$$\sum_{i=0}^{p} b_i = 1, \sum_{i=1}^{p} b_i = 1 - b_0,$$

于是

$$\beta_{1j} = \frac{b_j(1 - b_j)(1 + 2Ab_j)/(A + 1)}{\left[\frac{\sum_{1}^{p} b_i(1 - b_i)}{A + 1}\right]^2}$$

$$= \frac{b_j(1 - b_j)(1 + 2Ab_j)(A + 1)}{\left(\sum_{i=1}^{p} b_i(1 - b_i)\right)^2}, j = 1,2,\cdots,p,$$

$$\beta_{2j} = -\frac{\sum_{i=1}^{p} b_i(1 - b_i)}{\left[\sum_{i=1}^{p} b_i(1 - b_i)\Big/(A + 1)\right]^2}$$

$$= -\frac{A + 1}{\sum_{i=1}^{p} b_i(1 - b_i)}, j = 1,2,\cdots,p.$$

此时,相应的 Ω 矩阵是 $2p \times 2p$,可以分四块写出,即

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

我们引入 $\boldsymbol{b}$ 表示向量$(b_1, b_2, \cdots, b_p)'$,于是

$$\Omega_{11} = (D(\boldsymbol{b}) - \boldsymbol{bb}')/(A+1),$$
$$\Omega_{12} = (\mathrm{Cov}(x_i, x_j^2)) = \Omega'_{21}.$$

这在求 $\hat{A}_i$时已给出 $\mathrm{Cov}(x_i, x_i^2)$，而

$$\Omega_{22} = (\mathrm{Cov}(x_i^2, x_j^2)).$$

因此还需给出 $\mathrm{Cov}(x_i, x_j^2)$与 $\mathrm{Cov}(x_i^2, x_j^2)$. 从狄氏分布的性质，立即可得：当 $i \neq j$ 时

$$\begin{aligned}
\mathrm{Cov}(x_i, x_j^2) &= Ex_i x_j^2 - Ex_i Ex_j^2 \\
&= \frac{a_i a_j(a_j+1)}{A(A+1)(A+2)} - \frac{a_i}{A}\frac{a_j(a_j+1)}{A(A+1)} \\
&= -\frac{2a_i a_j(a_j+1)}{A^2(A+1)(A+2)} \\
&= -\frac{2b_i b_j(1+Ab_j)}{(A+1)(A+2)}, \\
\mathrm{Cov}(x_i^2, x_j^2) &= Ex_i^2 x_j^2 - Ex_i^2 Ex_j^2 \\
&= -\frac{2b_i b_j(1+Ab_i)(1+Ab_j)(2A+3)}{(A+1)^2(A+2)(A+3)},
\end{aligned}$$

而

$$\begin{aligned}
\mathrm{Var}(x_i^2) &= Ex_i^4 - (Ex_i^2)^2 \\
&= \frac{a_i(a_i+1)(a_i+2)(a_i+3)}{A(A+1)(A+2)(A+3)} - \left(\frac{a_i(a_i+1)}{A(A+1)}\right)^2 \\
&= \frac{b_i(1+b_i)[6A(A+1+b_i+Ab_i^2)+4A^3 b_i(1-Ab_i)]}{(A+1)^2(A+2)(A+3)}.
\end{aligned}$$

这样就求出

$$\begin{aligned}
\beta'\Omega\beta &= (\beta'_{(1)}\ \beta'_{(2)})\begin{pmatrix}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22}\end{pmatrix}\begin{pmatrix}\beta_{(1)} \\ \beta_{(2)}\end{pmatrix} \\
&= \beta'_{(1)}\Omega_{11}\beta'_{(1)} + 2\beta'_{(1)}\Omega_{12}\beta_{(2)} + \beta'_{(2)}\Omega_{22}\beta_{(2)} \\
&\overset{\text{记}}{=} c_{11} + 2c_{12} + c_{22}.
\end{aligned}$$

令

$$C_{11} = \frac{(A+1)^2}{\left(\sum_{i=1}^{p} b_i(1-b_i)\right)^4} \begin{pmatrix} b_1(1-b_1)(1+2Ab_1) \\ \vdots \\ b_p(1-b_p)(1+2Ab_p) \end{pmatrix}'$$

$$\times \{[D(\boldsymbol{b}) - \boldsymbol{bb}']/(A+1)\} \begin{pmatrix} b_1(1-b_1)(1+2Ab_1) \\ \vdots \\ b_p(1-b_p)(1+2Ab_p) \end{pmatrix}$$

$$= \frac{(A+1)}{\left(\sum_{i=1}^{p} b_i(1-b_i)\right)^4} \sum_{i=1}^{p} \sum_{j=1}^{p} b_i(1-b_i)(1+2Ab_i)$$

$$\times (\delta_{ij} b_i - b_i b_j) - b_j(1-b_j)(1+2Ab_j)$$

$$= \frac{(A+1)}{\left(\sum_{i=1}^{p} b_i(1-b_i)\right)^4} \left[\sum_{i=1}^{p} b_i^3(1-b_i)^2(1+2Ab_i)^2 \right.$$

$$\left. + \left(\sum_{i=1}^{p} b_i^2(1-b_i)(1+2Ab_i) \right)^2 \right] .$$

$$C_{12} = \beta'_{(1)} \Omega_{12} \beta_{(2)}$$

$$= -\frac{(A+1)^2}{\left(\sum_{i=1}^{p} b_i(1-b_i)\right)^3} \begin{pmatrix} b_1(1-b_1)(1+2Ab_1) \\ \vdots \\ b_p(1-b_p)(1+2Ab_p) \end{pmatrix}'$$

$$\times \left[-\left(\frac{\partial b_i b_j (1+Ab_j)}{(A+1)(A+2)} \right)_{ij=1,2,\cdots,p} \right] \mathbb{1}$$

$$= \frac{A+1}{(A+2)\left(\sum_{i=1}^{p} b_i(1-b_i)\right)^3} \left[2\left(\sum_{i=1}^{p} b_i^2(1-b_i)(1+2Ab_i) \right) \right.$$

$$\left. \times \left(\sum_{j=1}^{p} b_j(1+Ab_j) \right) \right]$$

$$= \frac{2(A+1)\left(\sum_{i=1}^{p} b_i^2(1-b_i)(1+2Ab_i) \right) \left(\sum_{i=1}^{p} b^i(1+Ab_i) \right)}{(A+2)\left(\sum_{i=1}^{p} b_i(1-b_i) \right)^3} .$$

同样地可以算出

$$C_{22}=\beta'_{(2)}\Omega_{22}\beta_{(2)}$$

$$=\frac{(A+1)^2}{\left(\sum_{i=1}^{p}b_i(1-b_i)\right)^2}\mathbb{1}'\Omega_{22}\mathbb{1}$$

$$=\frac{(A+1)^2}{\left(\sum_{i=1}^{p}b_i(1-b_i)\right)^2}\left[\frac{2\sum_{i=1}^{p}a_i(a_i+1)(2a_i+3)}{A(A+1)(A+2)(A+3)}\right.$$

$$\left.-\frac{2(2A+3)\left[\sum_{i=1}^{p}b_i(1+Ab_i)\right]^2}{(A+1)^2(A+2)(A+3)}\right]$$

$$=\frac{2(A+1)\sum_{i=1}^{p}b_i(1+Ab_i)(3+2Ab_i)-2(2A+3)\left(\sum_{i=1}^{p}b_i(1+Ab_i)\right)^2}{(A+2)(A+3)\left[\sum_{i=1}^{p}b_i(1-b_i)\right)^2}.$$

§3. 回　　归

现在考虑狄氏分布一部分变量对另一部分的条件期望与条件协方差矩阵，从而导出成分向量的分量与分量的回归形式.

设 $x=\begin{pmatrix}x_{(1)}\\x_{(2)}\end{pmatrix}=(x_0,x_1,\cdots,x_p)'$ 是一成分向量，$x_{(1)}=(x_0,x_1,\cdots,x_k)'$，$x_{(2)}=(x_{p+1},x_{k+2},\cdots,x_p)'$. 令 $x_0,x_1,\cdots,x_p$ 的联合密度是

$$\frac{\Gamma\left(\sum_{i=0}^{p}a_i\right)}{\prod_{i=0}^{p}\Gamma(a_i)}\prod_{i=0}^{p}x_i^{a_i-1},$$

$$x_0=1-\sum_{i=1}^{p}x_i>0,x_i>0,i=1,2,\cdots,p,$$

而 $x_{(1)}=(x_0,x_1,\cdots,x_k)'$的联合密度是

$$\frac{\Gamma\left(\sum_{i=0}^{p}a_i\right)}{\Gamma\left(\sum_{i=k+1}^{p}a_i\right)\prod_{i=0}^{k}\Gamma(a_i)}\left(\prod_{i=0}^{k}x_i^{a_i-1}\right)$$

$$\times\left(1-\sum_{i=0}^{k}x_i\right)^{\sum_{i=k+1}^{p}a_i-1},$$

$$0<x_i<1,i=0,1,\cdots,k,\sum_{i=0}^{k}x_i<1.$$

因此 $x_{(2)}$对 $x_{(1)}$的条件密度是

$$p(x_{(2)}\mid x_{(1)})$$

$$=\frac{\Gamma\left(\sum_{i=k+1}^{p}a_i\right)\prod_{i=k+1}^{p}x_i^{a_i-1}}{\left(\prod_{i=k+1}^{p}\Gamma(a_i)\right)\left(1-\sum_{i=0}^{k}x_i\right)^{\sum_{i=k+1}^{p}a_i-1}},$$

$$x_i>0,i=k+1,\cdots,p,\sum_{i=k+1}^{p}x_i<1-\sum_{i=0}^{k}x_i. \quad (3.1)$$

因此可以求得条件期望与条件方差，利用公式

$$\int_0^A x^{a-1}(A-x)^{b-1}\mathrm{d}x=A^{a+b-1}B(a,b)$$

就可求得

$$E\{x_i\mid x_{(1)}\}=\frac{a_i}{\sum_{j=k+1}^{p}a_j}\left(1-\sum_{j=0}^{k}x_j\right),$$

$$i=k+1,\cdots,p, \quad (3.2)$$

$$\mathrm{Var}\{x_i\mid x_{(1)}\}=\frac{a_i(a_{(k)}-a_i)}{a_{(k)}^2(a_{(k)}+1)}\left(1-\sum_{j=0}^{k}x_j\right)^2, \quad (3.3)$$

其中

$$a_{(k)}=\sum_{j=k+1}^{p}a_j,i=k+1,\cdots,p,$$

$$\mathrm{Cov}\{x_i, x_j \mid x_{(1)}\} = -\frac{a_i a_j}{a_{(k)}^2 (a_{(k)} + 1)} \left(1 - \sum_{t=0}^{k} x_t\right)^2,$$

$$i \neq j, i, j = k+1, \cdots, p. \tag{3.4}$$

上面求得的条件期望，对于考虑回归函数的形式是有帮助的.(3.2)式右端是两项的乘积，每一项的统计意义十分明显.条件期望是 $x_{(1)}$ 各分量的线性函数，而且系数都是相同的.

(3.2)式的一个特殊情形是 $k = p-1$，此时，(3.2)式就成为

$$E\{x_p \mid x_0, x_1, \cdots, x_{p-1}\} = 1 - \sum_{j=0}^{p-1} x_j,$$

由于 $\sum_{j=0}^{p} x_j = 1$，因此上式自然是成立的.然而，(3.2) 式给我们提供了另一种估计参数 $a_0, a_1, \cdots, a_p$ 的方法，就是利用成分向量的一部分对另一部分的回归，可以给出参数的估计，而方法与前面讨论过的是类似的.为了便于对比，将(3.2)、(3.3)、(3.4) 用另一个方式表示，就可以很方便看出相似性.令

$$y_i = x_i \Big/ \left(1 - \sum_{j=0}^{k} x_j\right), i = k+1, \cdots, p,$$

$$y_{(2)} = \begin{pmatrix} y_{k+1} \\ \vdots \\ y_p \end{pmatrix},$$

于是有

$$\begin{cases} E\{y_{(2)} \mid x_{(1)}\} = \dfrac{1}{a_{(k)}} \begin{pmatrix} a_{k+1} \\ \vdots \\ a_p \end{pmatrix} \xlongequal{\text{记}} b_{(k)}, \\ \mathrm{Var}\{y_{(2)} \mid x_{(1)}\} = \dfrac{1}{a_{(k)} + 1} [D(b_{(k)}) - b_{(k)} b'_{(k)}]. \end{cases} \tag{3.5}$$

将(3.5)中的条件期望看成是期望，(3.5)式与狄氏分布的

$$Ex = b, \mathrm{Var}(x) = \frac{1}{A+1}[D(b) - bb']$$

是完全相似的，$\mathbb{1}' b_{(k)} = 1$，这与$\mathbb{1}' b = 1$ 相似，参数 $a_{(k)}$与 A 完全相

似.所以前面讨论的用来估计 b 的方法,估计 A 的方法,均可用来估计 $b_{(k)}$和 $a_{(k)}$,这里就不再一一重复了.

更值得注意的是,条件期望的这一特性,在一定意义上是狄氏分布所独有的.在参考文献[4]中,有下述定理:

定理 3.1 设 $x_0,x_1,\cdots,x_p$ 是成分随机向量,即 $x_i>0,i=0,1,\cdots,p,\sum_{i=0}^{p}x_i=1$,有连续的密度函数 $p(x_1,\cdots,x_p)$,若 $p(x_1,\cdots,x_p)$ 满足:

(i) 在单形 D_p 上,$p(x_1,\cdots,x_p)>0$;

(ii) $p(x_1,\cdots,x_p)=f\left(\sum_{i=1}^{p}x_i\right)\prod_{i=1}^{p}x_i^{a_i-1}$,且 a_i 均大于 0;

(iii) 对任一正整数 $k,i=1,2,\cdots,p$ 使

$$E\left\{x_i^k\,\Big|\,\sum_{j\neq i}x_j\right\}=C(1-\sum_{j\neq i}x_j)^k,$$

其中 C 是一个常数,则$(x_0,x_1,\cdots,x_p)$一定是狄氏分布.

这一定理还可以进一步推广,将(ii)改为

$$p(x_1,\cdots,x_p)=f_0\left(\sum_{j=1}^{p}x_i\right)\prod_{i=1}^{p}f_i(x_i)$$

其中至少有一个 f_i 是幂函数.(iii)可改为

$$E\left\{x_i\,\Big|\,\sum_{j\neq i}x_j\right\}$$

是 $\sum_{j\neq i}x_j$ 的线性函数,定理的结论仍然成立.所以条件期望的这一特性,对于狄氏分布是具有特殊重要的意义.

§4. 判别分析

现在来求狄氏分布的判别函数.设有两个狄氏分布 $D(a_0,a_1,\cdots,a_p)$与 $D(b_0,b_1,\cdots,b_p)$,分别记为 $\boldsymbol{D}(a)$与 $\boldsymbol{D}(b)$,如果来自 $\boldsymbol{D}(a)$的概率为 π_a,来自 $\boldsymbol{D}(b)$的概率为 $\pi_b,\pi_a+\pi_b=1$.于是观察到样本 $x=(x_0,x_1,\cdots,x_p)'$后,相应的后验概率为

$$h(\boldsymbol{D}(a)\mid x)=\frac{\pi_a D(a_0,a_1,\cdots,a_p)}{\pi_a D(a_0,\cdots,a_p)+\pi_b D(b_0,\cdots,b_p)},$$

$$h(\boldsymbol{D}(b)\mid x)=\frac{\pi_b D(b_0,b_1,\cdots,b_p)}{\pi_a D(a_1,\cdots,a_p)+\pi_b D(b_1,\cdots,b_p)}.$$

当 $h(\boldsymbol{D}(a)\mid x)\geqslant h(\boldsymbol{D}(b)\mid x)$时，就将 x 判为来自 $\boldsymbol{D}(a)$；当 $h(\boldsymbol{D}(a)\mid x)\leqslant h(\boldsymbol{D}(b)\mid x)$时，将 x 判为来自 $\boldsymbol{D}(b)$. 因此 x 来自

$$\begin{aligned}\boldsymbol{D}(a)&\Leftrightarrow h(\boldsymbol{D}(a)\mid x)\geqslant h(\boldsymbol{D}(b)\mid x)\\&\Leftrightarrow \pi_a D(a_0,\cdots,a_p)\geqslant \pi_b D(b_0,\cdots,b_p)\\&\Leftrightarrow \frac{D(a_0,\cdots,a_p)}{D(b_0,\cdots,b_p)}\geqslant\frac{\pi_b}{\pi_a}.\end{aligned}$$

上式就是密度之比大于一个给定的值，而两个密度之比也可改写为两个密度各自取对数之后的差，这就导出

$$\begin{aligned}&\ln D(a_0,\cdots,a_p)-\ln D(b_0,\cdots,b_p)\\&=\ln\Gamma\Big(\sum_{i=0}^{p}a_i\Big)+\sum_{i=0}^{p}\big[(a_i-1)\ln x_i-\ln\Gamma(a_i)\big]\\&\quad-\ln\Gamma\Big(\sum_{i=0}^{p}b_i\Big)-\sum_{i=0}^{p}\big[(b_i-1)\ln x_i-\ln\Gamma(b_i)\big]\\&=\sum_{i=0}^{p}(a_i-b_i)\ln x_i+\ln\frac{\Gamma\Big(\sum_{i=0}^{p}a_i\Big)\prod_{i=0}^{p}\Gamma(b_i)}{\Gamma\Big(\sum_{i=0}^{p}b_i\Big)\prod_{i=0}^{p}\Gamma(a_i)},\end{aligned}$$

可见判别函数是 $\ln x_i$ 的线性函数，其中参数 a_i,b_i 可以从训练样本产生的估计量得到.

现在从判别信息量来考虑，如何导出相应的判别函数，注意到

$$I(D(a),D(b))=\mathrm{E}_a(\ln D(a_0,\cdots,a_p)-\ln D(b_0,\cdots,b_p)),$$

其中 E_a 表示对分布 $D(a_0,a_1,\cdots,a_p)$求期望值，因此利用上面的公式得到

$$I(D(a),D(b))=\sum_{i=0}^{p}(a_i-b_i)\mathrm{E}_a\ln x_i$$

$$+\ln\left[\frac{\Gamma\left(\sum_{i=0}^{p} a_i\right)}{\Gamma\left(\sum_{i=0}^{p} b_i\right)}\right]\prod_{i=0}^{p}\frac{\Gamma(b_i)}{\Gamma(a_i)}.$$

从狄氏分布密度导出的等式

$$\int\cdots\int_{D_p}\prod_{i=0}^{p} x_i^{a_i-1}\mathrm{d}x_1\cdots\mathrm{d}x_p=\frac{\prod_{i=0}^{p}\Gamma(a_i)}{\Gamma\left(\sum_{i=0}^{p}a_i\right)},$$

两边对 a_i 求偏微商,就得到

$$E_a\ln x_i=\frac{\Gamma'(a_i)}{\Gamma(a_i)}-\frac{\Gamma'(A)}{\Gamma(A)},$$

$$\left(A=\sum_{j=0}^{p}a_i\right)$$

代入 $I(\boldsymbol{D}(a),\boldsymbol{D}(b))$的表达式,就得到

$$I(\boldsymbol{D}(a),\boldsymbol{D}(b))=\sum_{i=0}^{p}(a_i-b_i)\left[\frac{\Gamma'(a_i)}{\Gamma(a_i)}-\frac{\Gamma'(A)}{\Gamma(A)}\right]+\ln\frac{\Gamma(A)}{\Gamma(B)}\prod_{i=0}^{p}\frac{\Gamma(b_i)}{\Gamma(a_i)},\tag{4.1}$$

其中 $B=\sum_{i=0}^{p}b_i$.

类似地,可以导出

$$I(\boldsymbol{D}(b),\boldsymbol{D}(a))=\sum_{i=0}^{p}(b_i-a_i)\left[\frac{\Gamma'(b_i)}{\Gamma(b_i)}-\frac{\Gamma'(B)}{\Gamma(B)}\right]+\ln\frac{\Gamma(B)}{\Gamma(A)}\prod_{i=0}^{p}\frac{\Gamma(a_i)}{\Gamma(b_i)}.\tag{4.2}$$

现在给了一个观察到的成分向量 $x=(x_0,\cdots,x_p)'$,当我们使用判别信息量时,如何去导出相应的判别函数呢?

由于判别信息量 $I(\boldsymbol{D}(a),\boldsymbol{D}(b))$是以 $\boldsymbol{D}(a)$为基准,衡量 $\boldsymbol{D}(b)$与它的差距.因此如果 $\boldsymbol{D}(a)$的参数是由训练样本给出的估计,记为 $\hat{a}=(\hat{a}_0,\hat{a}_1,\cdots,\hat{a}_p)'$.于是在新增了一个观察向量 x 之

后，就得另一个估计 $\hat{a}_*$. 将 $D(\hat{a}_*)$ 看成 $I(D(a),D(b))$ 中的 $D(b)$，这样 $I(D(\hat{a}),D(\hat{a}_*))$ 就给出了观察到的成分向量 x 与 $D(\hat{a})$ 的差距. 同样的方法，可以度量出 x 与另一个分布 $D(\hat{b})$ 的差距，哪一个差距小，就应归入哪一类分布，得到了判别的公式.

这一种考虑，同样地也可以用于对称的、由判别信息量导出的 $J(D(a),D(b))$，利用 J 与 I 的关系，就得到

$$J(D(a),D(b)) = \sum_{i=0}^{p}(a_i - b_i)\left[\frac{\Gamma'(a_i)}{\Gamma(a_i)} - \frac{\Gamma'(b_i)}{\Gamma(b_i)} + \frac{\Gamma'(B)}{\Gamma(B)} - \frac{\Gamma'(A)}{\Gamma(A)}\right]. \quad (4.3)$$

此时，用训练样本算出 a 的估计 $\hat{a}$，用训练样本添加新的观察数据 x 算出 a 的估计 $\hat{a}_*$，用 $\hat{a}$ 代 a，$\hat{a}_*$ 代 b 就可以求出 x 与 $D(\hat{a})$ 的差距. 由于 $\hat{a}$ 的表达式不同，有的还无法求出表达式(如最大似然估计)，所以在实际中应用(4.3)式也只能给出一种算法. 为了便于说明，我们用一个比较特殊的例子来解释它的用法.

假定从 $D(a)$ 中已获得的训练样本是 $\underset{n\times(p+1)}{X}$，它的均值向量是 $\underset{(p+1)\times 1}{\bar{x}} = \frac{1}{n}X'\mathbb{1}$. 注意到(4.3)可以改写为

$$J(D(a),D(b)) = A\sum_{i=0}^{p}\left(\frac{a_i}{A} - \frac{B}{A}\frac{b_i}{B}\right)\left[\frac{\Gamma'(a_i)}{\Gamma(a_i)} - \frac{\Gamma'(A)}{\Gamma(A)} + \frac{\Gamma'(B)}{\Gamma(B)} - \frac{\Gamma'(b_i)}{\Gamma(b_i)}\right].$$

这样上式右端的项，分别用各自的估计量代入，注意到似然方程给出的估计，就可以用 X 中的元素 $x_{\alpha i}$ 和得到的观察值 x 来估计 $J(D(a),D(b))$，从而得到 x 与 $D(a)$ 的距离. 因为

$$\frac{1}{n}\sum_{\alpha=1}^{n}\ln x_{\alpha i} \text{ 可以估计} \frac{\Gamma'(a_i)}{\Gamma(a_i)} - \frac{\Gamma'(A)}{\Gamma(A)}, \text{(似然方程)}$$

$$\frac{1}{n}\sum_{\alpha=1}^{n}x_{\alpha i} \text{ 可以估计} \frac{a_i}{A}, \text{(矩估计)}$$

$$\frac{\sum_{i=1}^{p} \bar{x}_i(1-\bar{x}_i)}{\sum_{i=1}^{p} S_i^2} - 1 \text{ 可以估计 } A,\text{(线性模型的估计)}$$

其中

$$\bar{x}_i = \frac{1}{n}\sum_{\alpha=1}^{n} x_{\alpha i}, S_i^2 = \frac{1}{n-1}\sum_{\alpha=1}^{n}(x_{\alpha i} - \bar{x}_i)^2,$$

添加了一个新观察到的 $x=(x_0,\cdots,x_p)$之后，

$$\frac{1}{n+1}\left(\sum_{\alpha=1}^{n}\ln x_{\alpha i} + \ln x_i\right)\text{可以估计}\frac{\Gamma'(b_i)}{\Gamma(b_i)} - \frac{\Gamma'(B)}{\Gamma(B)},$$

$$\frac{1}{n+1}\left(\sum_{\alpha=1}^{n} x_{\alpha i} + x_i\right)\text{可以估计}\frac{b_i}{B},$$

$$\frac{\sum_{i=1}^{p} \bar{x}_{i*}(1-\bar{x}_{i*})}{\sum_{i=1}^{p} S_{i*}^2} - 1 \text{ 可以估计 } B,$$

其中

$$\bar{x}_{i*} = \frac{1}{n+1}\left(\sum_{\alpha=1}^{n} x_{\alpha i} + x_i\right)$$

$$S_{i*}^2 = \frac{1}{n}\left(\sum_{\alpha=1}^{n}(x_{\alpha i} - \bar{x}_{i*})^2 + (x_i - \bar{x}_{i*})^2\right).$$

于是就得到 $J(D(a),D(b))$的估计值是

$$\hat{J} = \sum_{i=0}^{p}\left[\bar{x}_i - \left(\frac{1}{n+1}(n\bar{x}_i + x_i)\right)\left(\frac{\hat{B}}{\hat{A}}\right)\right]$$

$$\times\left[\frac{1}{n}\sum_{\alpha=1}^{n}\ln x_{\alpha i} - \frac{1}{n+1}\left(\sum_{\alpha=1}^{n}\ln x_{\alpha i} + \ln x_i\right)\right]\hat{A}$$

$$= \sum_{i=0}^{p}(\hat{A}\bar{x}_i - \hat{B}(n\bar{x}_i + x_i)/(n+1))\frac{\sum_{\alpha=1}^{n}\ln x_{\alpha i}/x_i}{(n+1)n}$$

$$= \frac{1}{n(n+1)^2}\sum_{i=0}^{p}[((n+1)\hat{A} - n\hat{B})\bar{x}_i - \hat{B}x_i]$$

$$\times \left(\sum_{\alpha=1}^{n} \ln x_{\alpha i}/x_i \right),$$

其中

$$\hat{A} = \sum_{i=1}^{p} \bar{x}_i (1 - \bar{x}_i) \Big/ \sum_{i=1}^{p} S_i^2,$$

$$\hat{B} = \sum_{i=1}^{p} \bar{x}_{i*} (1 - \bar{x}_{i*}) \Big/ \sum_{i=1}^{p} S_{i*}^2.$$

这样就求出了 $x=(x_0,\cdots,x_p)'$ 对训练样本 $\underset{n\times(p+1)}{X}$ 所形成的分布 $\boldsymbol{D}(a)$ 的距离 $\hat{J}$. 同样的方法和相似的表达式，就可以给出 x 对另一个训练样本给出的分布有多大距离，比较距离大小，就可以决定 x 属于哪一个分布，达到了判别的要求.

§5. 典型相关分析

狄氏分布的方差协方差矩阵有特殊的结构，它与分布 $D(a_0,\cdots,a_p)$ 中的参数有密切的联系，因此成分向量的部分分量之间的典型相关的分析方法有一些特殊性.

我们从 $(x_0,\cdots,x_p)$ 成分向量的方差协方差阵来分析. 今

$$V(x) = V(x_0,\cdots,x_p) = \frac{1}{A+1}(D(b) - bb'),$$

将 x 分为两部分 $x=(x'_{(1)} \quad x'_{(2)})'$ 后，$V(x)$ 和 b 也相应分块，得

$$V(x) = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \begin{matrix} k+1 \\ p-k. \end{matrix}$$

注意到 b 相应分为 $(b'_{(1)}\ b'_{(2)})'$ 后，就有

$$V_{ii} = \frac{1}{A+1}(D(b_{(i)}) - b_{(i)}b'_{(i)}), i = 1,2,$$

$$V_{ij} = \frac{-1}{A+1} b_{(i)} b'_{(j)}, \qquad i \neq j, i,j = 1,2.$$

而由于 $b'_{(1)}\mathbb{1} \neq 1, i=1,2$. 因此

$$V_{ii}^{-1} = (A+1)\left(D(b_{(i)}^{-1}) + \frac{\mathbb{1}\mathbb{1}'}{1 - b'_{(i)}\mathbb{1}} \right), i = 1,2,$$

其中

$$b_{(1)}^{-1} = (b_0^{-1}, \cdots, b_k^{-1})',$$

$$b_{(2)}^{-1} = (b_{k+1}^{-1}, \cdots, b_p^{-1})'.$$

通常情况下,求典型相关系数,应考虑矩阵

$$V_{11}^{-1} V_{12} V_{22}^{-1} V_{21}$$

的特征根,它有多少个非零特征根,就有多少对典型相关的变量.但对狄氏分布这一特殊情况,注意到

$$V_{12} = -\frac{1}{A+1} b_{(1)} b'_{(2)},$$

它的秩是 1,因此 $V_{11}^{-1} V_{12} V_{22}^{-1} V_{21}$ 有而且只有一个非零特征向量,所以只有一对典型相关变量.由于成分向量分段为 $x_{(1)}$ 与 $x_{(2)}$ 后,有

$$1 = \mathbb{1}'x = \mathbb{1}'x_{(1)} + \mathbb{1}'x_{(2)},$$

明显地表明了,$\mathbb{1}'x_{(1)}$ 与 $\mathbb{1}'x_{(2)}$ 是一对典型相关变量.由于

$$\begin{aligned} \mathrm{Var}(\mathbb{1}'x_{(1)}) &= \mathbb{1}'(D(b_{(1)}) - b_{(1)} b'_{(1)}) \mathbb{1}/(A+1) \\ &= \mathbb{1}'b_{(1)}(1 - \mathbb{1}'b_{(1)})/(A+1), \\ \mathrm{Var}(\mathbb{1}'x_{(2)}) &= \mathbb{1}'b_{(2)}(1 - \mathbb{1}'b_{(2)})/(A+1), \\ \mathrm{Cov}(\mathbb{1}'x_{(1)}, \mathbb{1}'x_{(2)}) &= -(\mathbb{1}'b_{(1)})(\mathbb{1}'b_{(2)})/(A+1). \end{aligned}$$

因此取

$$\xi = \sqrt{\frac{A+1}{\mathbb{1}'b_{(1)} b'_{(2)} \mathbb{1}}}\, \mathbb{1}_{k+1},$$

$$\eta = \sqrt{\frac{A+1}{\mathbb{1}'b_{(1)} b'_{(2)} \mathbb{1}}}\, \mathbb{1}_{p-k}$$

之后,由于 $1 = \mathbb{1}'b = \mathbb{1}'b_{(1)} + \mathbb{1}'b_{(2)}$,就有

$$\mathrm{Var}(\xi' x_{(1)}) = 1, \mathrm{Var}(\eta' x_{(2)}) = 1,$$

$$\mathrm{Cov}(\xi' x_{(1)}, \eta' x_{(2)}) = -1.$$

因此 $\xi' x_{(1)}$ 与 $\eta' x_{(2)}$ 是一对典型变量,其余与 ξ 正交的向量作为系数的 $x_{(1)}$ 的线性函数,与 η 正交的向量作为系数的 $x_{(2)}$ 的线性函数,彼此都是不相关的,协方差为 0.这一点,狄氏分布与逻辑正态

是很不相同的.从上面的讨论可以看出,凡是与 $b_{(1)}$ 正交的向量,也即对向量

$$\alpha = \left(I_{k+1} - \frac{1}{\| b_{(1)} \|} b_{(1)} b'_{(1)}\right) u, u \text{ 任意},$$

总有

$$\begin{aligned}
&\mathrm{Cov}(\alpha' x_{(1)}, x_{(2)}) \\
&= -\frac{1}{A+1} u' \left(I - \frac{b_{(1)} b'_{(1)}}{\| b_{(1)} \|}\right) b_{(1)} b'_{(2)} = 0.
\end{aligned}$$

同样地,对

$$\beta = \left(I_{p-k} - \frac{1}{\| b_{(2)} \|} b_{(2)} b'_{(2)}\right) v, v \text{ 任意},$$

也总有

$$\begin{aligned}
&\mathrm{Cov}(x_{(1)}, \beta' x_{(2)}) \\
&= -b_{(1)} b'_{(2)} \left(I_{p-k} - \frac{b_{(2)} b'_{(2)}}{\| b_{(2)} \|}\right) \Big/ (A+1) = 0,
\end{aligned}$$

适当选择 u 与 v,使 $\alpha' V_{11} \alpha = 1, \beta' V_{22} \beta = 1$,就找到了一对不相关的典型变量 $\alpha' x_{(1)}$ 与 $\beta' x_{(2)}$.注意到如此选择的 α, β 可以与所需的成对典型变量的数目相符,这样就找出了全部的典型相关变量.

现在引入样本矩阵 $\underset{n \times (p+1)}{X}$,对于样本矩阵 X,相应的样本协方差阵是

$$\begin{aligned}
S &= \frac{1}{n-1} L_{xx} = \frac{1}{n-1} X' (I - \frac{1}{n} \mathbb{1} \mathbb{1}') X \\
&= \frac{1}{n-1} (X'X - n \bar{x} \bar{x}'),
\end{aligned}$$

其中 $\bar{x} = \frac{1}{n} X' \mathbb{1}$, $\bar{x}$ 是 b 的无偏估计.

将成分向量 x 分成两段:$x = (x'_{(1)}, x'_{(2)})'$,相应的 X 也分成两个子阵,即

$$X = (X_1 \quad X_2),$$

同样地 $\bar{x}$ 也分成相应的 $\bar{x}_{(1)}, \bar{x}_{(2)}$,于是

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{matrix} k+1 \\ p-k \end{matrix},$$

$$(n-1)S_{11} = X_1'X_1 - n\bar{x}_{(1)}\bar{x}'_{(1)},$$

$$(n-1)S_{22} = X_2'X_2 - n\bar{x}_{(2)}\bar{x}'_{(2)},$$

$$(n-1)S_{12} = X_1'X_2 - n\bar{x}_{(1)}\bar{x}'_{(2)}.$$

但样本协差阵并不像总体分布所对应的协差阵，S_{12}的秩会大于1.然而我们知道应该只有一个非零的奇异值，所以只需考虑 S_{12} 的最大的奇异值就行了.从前面的讨论知道，这个奇异值就是 1，因为只有一对典型变量，相关系数是 -1(也可以是 1，只需将一个系数向量改号就行).因此，利用这一特性，也就可以给出对狄氏分布参数估计的方法.

§6. 贝叶斯方法

对狄氏分布的讨论，从上面的几节已经可以看到是有难度的.就是用贝叶斯方法，也是不像正态、指数分布那样，可以有一些明确的表达式，主要困难在于后验分布的形式很复杂，难以计算.这一节介绍的是用 MCMC 方法(马可夫链蒙特卡罗方法)得到的一些结论，详细的内容可参看文献[6].

尽管狄氏分布是指数族分布的一个成员，这从指数族的定义很快就可以得到，但它的共轭先验并不好处理.我们现在来看一下这个困难所在.

指数族的定义是这样的：若样本 x 的分布密度 $p(x;\theta)$具有下述(6.1)的形式，则称分布属于指数族分布，即要求

$$p(x;\theta) = C(\theta)g(x)e^{t'(x)R(\theta)}, \tag{6.1}$$

其中 θ 是未知参数，$t(x)$是样本 x 的函数，是统计量.

上式也可以有另一种表示法，即(6.1)可改写为

$$p(x;\theta) = g(x)e^{t'(x)R(\theta)-\psi(\theta)}, \tag{6.2}$$

只要取 $\psi(\theta) = \ln\dfrac{1}{C(\theta)}$就行了.

关于指数族分布，有许多好的性质，例如，$t(x)$是充分统计量，期望和方差协方差矩阵会有好的表达式等等，这里着重于它的共轭先验分布．对于(6.2)式，容易看出，如果我们选用先验分布有如下的形式：

$$\pi(\theta|\mu,\lambda)=K(\mu,\lambda)e^{\mu' R(\theta)-\lambda\psi(\theta)}, \tag{6.3}$$

那么，用贝叶斯公式，就得后验分布

$$h(\theta|x)=\frac{e^{(t(x)+\mu)'R(\theta)-(\lambda+1)\psi(\theta)}}{\int e^{(t(x)+\mu)'R(\theta)-(\lambda+1)\psi(\theta)}d\theta},$$

可见(6.3)与$h(\theta|x)$仍属于同一类型，也即有

$$h(\theta|x)=\pi(\theta|\mu+t(x),\lambda+1). \tag{6.4}$$

因此贝叶斯方法能否比较好用，与(6.3)式给出先验分布$\pi(\theta\ \mu,\lambda)$是否好处理就密切相连．

容易验证狄氏分布$D(a_0,a_1,\cdots,a_p)$是属于指数族分布的，因为它的密度

$$p(x_1,\cdots,x_p;a_0,\cdots,a_p)=\frac{\Gamma\left(\sum\limits_{i=0}^{p}a_i\right)}{\prod\limits_{i=0}^{p}\Gamma(a_i)}\prod_{i=0}^{p}x_i^{a_i-1}$$

$$=\left(\prod_{i=0}^{p}x_i^{-1}\right)e^{\sum\limits_{i=0}^{k}a_i\ln x_i-\left(\sum\limits_{i=0}^{p}\ln\Gamma(a_i)-\Gamma(A)\right)},$$

其中

$$A=\sum_{i=0}^{p}a_i, x_0=1-\sum_{i=1}^{p}x_i,$$

$$x_i>0,\sum_{i=1}^{p}x_i<1.$$

因此，相应于(6.2)中的

$$t'(x)=(\ln x_0,\ln x_1,\cdots,\ln x_p)',$$

$$R(\theta)=a=(a_0,a_1,\cdots,a_p)',$$

$$\psi(\theta)=\sum_{i=0}^{p}\Gamma(a_i)-\Gamma\left(\sum_{i=0}^{p}a_i\right),$$

参数

$$\theta = a = R(\theta).$$

因此共轭的先验分布是

$$\pi(a \mid \mu,\lambda) = K(\mu,\lambda)\mathrm{e}^{\sum_{i=0}^{p} a\mu_i - \lambda\psi(a)},$$

注意到 $\psi(a)$是一个很复杂的函数,所以实际上是很难处理的.

为了说明这一点,我们看一个最简单的情形,贝他分布的密度

$$p(x;\theta_1,\theta_2) = \frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} x^{\theta_1 - 1}(1 - x)^{\theta_2 - 1}, 0 < x < 1.$$

$$\pi(\theta_1,\theta_2 \mid \mu,\lambda) \propto x^{\mu_1}(1 - x)^{\mu_2}\mathrm{e}^{-\lambda\psi(\theta)}$$

其中

$$\psi(\theta) = \ln\Gamma(\theta_1) + \ln\Gamma(\theta_2) - \ln\Gamma(\theta_1 + \theta_2),$$

符号$\propto$表示两端只少一个与变量 θ_1,θ_2 无关的常数因子.即使认为 θ_1 已知

$$\pi(\theta_2 \mid \mu,\lambda) \propto \left(\frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_2)}\right)^{\lambda}, \mu = 0,$$

此时后验分布

$$h(\theta_2 \mid x,\lambda) \propto \left(\frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_2)}\right)^{\lambda+1}(1 - x)^{\theta_2},$$

这也是很不好处理的.

就是对无信息先验分布,因为 $a_0, a_1, \cdots, a_p$ 的变化范围是 $p+1$维空间中无界的正象限,所以也要作一些修改,才能算出一些结果.下面我们扼要地介绍一下 MCMC 方法,然后列出用这个算法求得的数字结果,不难看出,这一方法对于上面谈到的共轭先验分布是行得通的.

MCMC 方法通常有两种算法,这里用的是 Metropolies 算法.这个算法的要点是这样的:

对给定的密度 $p(u \mid x)$,选一个辅助的密度 $q(v,u)$,它对每个 u 是一个 v 的密度,并且它是 u,v 的对称函数,即有

$$q(u,v) = q(v,u).$$

于是可以有一个算法如下：

(1)记 u 的值在第 t 步开始时是 $u=u^{(t-1)}$，于是从 $q(v,u)$ 中取一个样本 v.

(2)计算 $\Gamma=p(v|x)/p(u^{(t-1)}|x)$.

(3)若 $\Gamma\geqslant 1$，取 $u^{(t)}=v$；若 $\Gamma<1$，则

$$u^{(t)}=\begin{cases} v,\text{按概率 } \Gamma \text{ 来取;} \\ u^{(t-1)},\text{按概率 } 1-\Gamma \text{ 来取}. \end{cases}$$

这样继续下去，实际上，当 $t\to\infty$ 时，$u^{(t)}$ 依分布收敛于密度为 $p(u|x)$ 的随机变量，而且 $p(u|x)$ 可以选 $L(u|x)\pi(u)$，$L(u|x)$ 是似然函数，$\pi(u)$ 是选定的先验分布.

对于狄氏分布 $D(a_1,a_2,\cdots,a_k)$，取

$$A=\sum_{j=1}^{k}a_j, M \text{ 是一个大的正数},$$

$$\pi(a_1,\cdots,a_k)\propto A^{k-\frac{1}{2}},\quad e^{-M}<A<e^{M}$$

于是

$$p(a_1,\cdots,a_k|x)\propto A^{k-\frac{1}{2}}\left\{\frac{\Gamma(A)}{\prod_{i=1}^{k}\Gamma(a_i)}\right\}^{n}\prod_{j=1}^{k}\left(\prod_{\alpha=1}^{n}x_{\alpha_j}\right)^{a_j-1}.$$

再将参数 $a_1,\cdots,a_k$ 作变换：

$$\theta_i=\operatorname{log}it(a_i/A), i=1,2,\cdots,k-1,$$

$$\theta_k=\ln A,\text{而 } \operatorname{log}it\, x=\ln\frac{x}{1-x}.$$

$q(v,u)$ 选正态分布，期望向量是 u，协方差矩阵选 θ 的信息阵估计值 $\sum_{\hat\theta}$ 乘以 c^2，而

$$\sum\nolimits_{\hat\theta}=\left[-\left(\frac{\partial^2\ln p(\theta|x)}{\partial\theta_i\partial\theta_j}\right)\bigg|_{\theta=\hat\theta}\right]^{-1},\qquad(6.5)$$

$c^2=2.4/\sqrt{k}$，这样选的理由是由别人提供的经验.

下面的例子是关于北京鸭各种蛋白质在血浆中的相对比例，

表 6.1　三种蛋白质的相对比例

x_1:胎蛋白　　x_2:蛋白　　x_3:球蛋白

x_1	x_2	x_3	x_1	x_2	x_3
0.178	0.346	0.476	0.060	0.435	0.505
0.162	0.307	0.531	0.089	0.418	0.493
0.083	0.448	0.469	0.050	0.485	0.465
0.087	0.474	0.439	0.073	0.378	0.549
0.078	0.503	0.419	0.064	0.562	0.374
0.040	0.456	0.504	0.085	0.465	0.450
0.049	0.363	0.588	0.094	0.388	0.518
0.100	0.317	0.583	0.014	0.449	0.537
0.075	0.394	0.531	0.060	0.544	0.396
0.084	0.445	0.471	0.031	0.569	0.400
□	□	□	0.025	0.491	0.484
□	□	□	0.045	0.613	0.342
□	□	□	0.0195	0.526	0.4545

最早是由文献[7]中给出的狄氏分布的最大似然估计值,现在在文献[5]中,用上述 MCMC 方法算出相应的值,以便比较.

此时 $A=a_1+a_2+a_3, \theta_1=\operatorname{log}it\,\frac{a_1}{A}, \theta_2=\operatorname{log}it\,\frac{a_2}{A}, \theta_3=\log A$,注意 $\operatorname{log}it\,x=\log\frac{x}{1-x}$.

将数据代入变换后的后验密度,求得后验密度的众数(mode)是

$$\hat{\theta}_1=-2.55713,\quad \hat{\theta}_2=-0.201658,\quad \hat{\theta}_3=3.78617.$$

并求得由(6.4)得

$$\hat{\sum}_{\hat{\theta}}=\begin{bmatrix}0.0133824 & * & * \\ -0.00173654 & 0.00385699 & * \\ -0.00611651 & 0.000732927 & 0.044459\end{bmatrix}.$$

相应的相关矩阵是

$$R=\begin{bmatrix}1 & -0.241441 & -0.250369 \\ * & 1 & 0.0547773 \\ * & * & 1\end{bmatrix}.$$

经过 800 次迭代，稳定后再进入下一阶段. 这样就可以用 MCMC 求得参数估计.

由于有先验信息知道 $0<Ex_1<\frac{1}{3}$，因此先验分布的选择是在

$$\left[0,\frac{1}{3}\right]\times[0,1]\times[e^{-10},e^{10}]$$

上的均匀分布. 现列表给出文献[7]中最大似然的结果，MCMC 给出的结果，以便比较.

表 6.2　最大似然的估计值

参　数	最大似然估计值	90%置信区间
a_1	3.22	(1.11,4.33)
a_2	20.38	(13.20,27.48)
a_3	21.68	(14.14,29.23)

MCMC 的结果可以求出分位数的值，所以对分位数作比较就更有意义.

表 6.3　参数估计的分位数表

参　数		后验分布 0.05 分位数点	后验分布 0.50 分位数点	后验分布 0.95 分位数点
文献[7]的结果	a_1	2.067	2.998	4.167
	a_2	12.68	18.86	26.09
	a_3	13.59	20.01	27.64
MCMC 的结果	a_1	2.071	3.005	4.158
	a_2	12.914	18.915	25.931
	a_3	13.720	20.174	27.902

从这些数值就可以看出，结果是非常接近的.

因此对 MCMC 方法，应该尽可能给以介绍，国内这一方面的材料还不多.

全书到此就结束了，在这一章中，提出了一些处理狄氏分布问题的想法，展示了一些计算结果，目的是说明它的难处，但也不是完全没有办法解决.

我们没有讨论假设检验问题，主要困难在于似然比的统计量给不出表达式，实际上也难以使用．用贝叶斯方法同样涉及后验分布的计算，必须详细介绍 MCMC 方法，这又扩大了本书的内容，因此只能作罢．

习　题　四

1．证明公式(2.7)．(注意习题三中 A^+ 的定义)

2．如何求出(2.7)的表达式．利用 $D(b)-bb'$ 的秩是 1，并且 $\mathbb{1}$ 是它的 0 特征值对应的特征向量，因此记 $A=D(b)-bb'$ 后，就有

$$AA^+ = I - \frac{1}{p+1}\mathbb{1}\mathbb{1}' = A^+A.$$

于是可以令

$$A^+ = D(b^{-1}) + a_1\mathbb{1}\mathbb{1}' + a_2(b\mathbb{1}' + \mathbb{1}b'),$$

然后代入求 a_1, a_2 的值．

3．求出 $\hat{A}_{(2)}, \hat{A}_{(3)}$ 的渐近方差．

4．对狄氏分布 $D(a_0, a_1, \cdots, a_p)$，统计量 $(\ln x_0, \ln x_1, \cdots, \ln x_p)$ 是 $(a_0, a_1, \cdots, a_p)$ 参数的充分统计量，它是不是完备的？

5．用 GMM 方法时，用信息阵作为度量的矩阵，能否写出有关的算法？

参　考　文　献

[1] Hamilton, J. D. (1994), Time Series Analysis, Princeton University Press.

[2] 张尧庭、方开泰(1982,1998)，多元统计分析引论，科学出版社．

[3] 张尧庭(1991)，定性资料的统计分析，广西师范大学出版社．

[4] Gupta, R. D. and Richards, D (1990), The Dirichlet distributions and polynomial regression, Journal of Multivariate analysis, 32, 95—102.

[5] 章栋恩(1999)，成分数据的几个估计(待发表)；Bayesian inference on compositional data sampling from Dirichlet distribution．(待发表)．

[6] Tanner, M. A. (1996), Tools for Statistical Inference (Third Edition), Springer.

[7] Mosimann, J. E. (1962), On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions, Biometrika, 49, 65—81.

汉英名词对照表

三画

大小 size (1.2)

子成分 subcomposition (1.2)

广义矩估计法 generalized moment estimate method (4.1)

马氏链蒙特卡罗方法,MCMC方法 Markov chain Monte-Carlo method (4.6)

四画

贝他分布 beta distribution (1.5)

方向性数据的分布 directional data distribution (2.5)

尺度 scale (2.5)

贝叶斯方法 Bayesian method (4.6)

五画

对比 contrast (1.2)

正态分布 normal (Gauss) distribution (1.3)

～多元正态 multivariate normal (1.3)

～标准正态 standard normal (1.3)

～反正态 inverse Gaussian (1.3)

～逆反正态 reciprocal inverse Gaussian (1.3)

～对数正态 lognormal (1.4)

正态密度 normal density (1.3)

正态分布的特征函数 characteristic function of normal distribution (1.3)

对数比 logratio (2.2)

对数对比 logcontrast (2.2)

主分量 principle component (3.3)

六画

成分 composition (1.2)

各向同性 isotropic covariance (2.2)

七画

形状 shape (1.2)

伽马分布 gamma distribution (1.5)

狄氏分布 Dirichlet distribution (1.5)

～逆狄氏分布 inverse Dirichlet distribution (1.5)

判别信息量 discriminate information (3.3)

八画

径可分分布 radially decomposable distribution (2.5)

线性模型 linear model (4.1)

九画

总量 total (1.2)

十一画

基 basis (1.2)

基本向量 base vector (2.5)

球对称 spherical symmetric (2.5)

球面上均匀 spherical uniform (2.5)